XLSTAT 2024

https://www.xlstat.com

Introduction

XLSTAT started in 1995 in order to make accessible to anyone a powerful, complete and user-friendly data analysis and statistical solution.

The **accessibility** comes from the compatibility of XLSTAT with all the Microsoft Excel versions that are used nowadays (starting from Excel 2003 up to Excel 2016), from the interface that is available in several languages (English, French, German, Italian, Japanese, Portuguese, Spanish, …) and from the permanent availability of a fully functional 30 days evaluation version on the XLSTAT website www.xlstat.com.

The **power** of XLSTAT comes from both the C++ programming language, and from the algorithms that are used. The algorithms are the result of many years of research of thousands of statisticians, mathematicians, computer scientists throughout the world. Each development of a new functionality in XLSTAT is preceded by an in-depth research phase that sometimes includes exchanges with the leading specialists of the methods of interest.

The **completeness** of XLSTAT is the fruit of over fifteen years of continuous work, and of regular exchanges with the users' community. Users' suggestions have helped a lot improving the software, by making it well adapted to a variety of requirements.

Last, the **usability** comes from the user-friendly interface, which after a few minutes of trying it out, facilitates the use of some statistical methods that might require hours of training with other software.

The software architecture has considerably evolved over the last 5 years in order to take into account the advances of Microsoft Excel and the compatibility issues between platforms. The software relies on Visual Basic Application for the interface and on C++ for the mathematical and statistical computations.

As always, the Addinsoft team and the XLSTAT distributors are available to answer any question you have, or to take into account your remarks and suggestions in order to continue improving the software.

Installation

License

XLSTAT 2018 - SOFTWARE LICENSE AGREEMENT

ADDINSOFT SARL ("ADDINSOFT") IS WILLING TO LICENSE VERSION 2018 OF ITS XLSTAT (r) SOFTWARE AND THE ACCOMPANYING DOCUMENTATION (THE "SOFTWARE") TO YOU ONLY ON THE CONDITION THAT YOU ACCEPT ALL OF THE TERMS IN THIS AGREEMENT. PLEASE READ THE TERMS CAREFULLY. BY USING THE SOFTWARE YOU ACKNOWLEDGE THAT YOU HAVE READ THIS AGREEMENT, UNDERSTAND IT AND AGREE TO BE BOUND BY ITS TERMS AND CONDITIONS. IF YOU DO NOT AGREE TO THESE TERMS, ADDINSOFT IS UNWILLING TO LICENSE THE SOFTWARE TO YOU.

1. LICENSE. Addinsoft hereby grants you a nonexclusive license to install and use the Software in machine-readable form on a single computer for use by a single individual if you are using the demo version or if you have registered your demo version to use it with no time limits. If you have ordered a multi- users license, the number of users depends directly on the terms specified on the invoice sent to your company by Addinsoft or the authorized reseller.

2. RESTRICTIONS. Addinsoft retains all right, title, and interest in and to the Software, and any rights not granted to you herein are reserved by Addinsoft. You may not reverse engineer, disassemble, decompile, or translate the Software, or otherwise attempt to derive the source code of the Software, except to the extent allowed under any applicable law. If applicable law permits such activities, any information so discovered must be promptly disclosed to Addinsoft and shall be deemed to be the confidential proprietary information of Addinsoft. Any attempt to transfer any of the rights, duties or obligations hereunder is void. You may not rent, lease, loan, or resell for profit the Software, or any part thereof. You may not reproduce or distribute the Software except as expressly permitted under Section 1, and you may not create derivative works of the Software unless with the express agreement of Addinsoft.

3. SUPPORT. Registered users of the Software are entitled to Addinsoft standard support services. Demo version users may contact Addinsoft for support but with no guarantee to benefit from Addinsoft standard support services.

4. NO WARRANTY. THE SOFTWARE IS PROVIDED "AS IS" AND WITHOUT ANY WARRANTY OR CONDITION, WHETHER EXPRESS, IMPLIED OR STATUTORY. Some jurisdictions do not allow the disclaimer of implied warranties, so the foregoing disclaimer may not apply to you. This warranty gives you specific legal rights and you may also have other legal rights which vary from state to state, or from country to country.

5. LIMITATION OF LIABILITY. IN NO EVENT WILL ADDINSOFT OR ITS SUPPLIERS BE LIABLE FOR ANY LOST PROFITS OR OTHER CONSEQUENTIAL, INCIDENTAL OR SPECIAL DAMAGES (HOWEVER ARISING, INCLUDING NEGLIGENCE) IN CONNECTION WITH THE SOFTWARE OR THIS AGREEMENT, EVEN IF ADDINSOFT HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. In no event will Addinsoft liability in connection with the Software, regardless of the form of action, exceed the price paid for acquiring the Software. Some jurisdictions do not allow the foregoing limitations of liability, so the foregoing limitations may not apply to you.

6. TERM AND TERMINATION. This Agreement shall continue until terminated. You may terminate the Agreement at any time by deleting all copies of the Software. This license terminates automatically if you violate any terms of the Agreement. Upon termination you must promptly delete all copies of the Software.

7. CONTRACTING PARTIES. If the Software is installed on computers owned by a corporation or other legal entity, then this Agreement is formed by and between Addinsoft and such entity. The individual executing this Agreement represents and warrants to Addinsoft that they have the authority to bind such entity to the terms and conditions of this Agreement.

8. INDEMNITY. You agree to defend and indemnify Addinsoft against all claims, losses, liabilities, damages, costs and expenses, including attorney's fees, which Addinsoft may incur in connection with your breach of this Agreement.

9. GENERAL. The Software is a "commercial item." This Agreement is governed and interpreted in accordance with the laws of the Court of Paris, France, without giving effect to its conflict of laws provisions. The United Nations Convention on Contracts for the International Sale of Goods is expressly disclaimed. Any claim arising out of or related to this Agreement must be brought exclusively in a court located in PARIS, FRANCE, and you consent to the jurisdiction of such courts. If any provision of this Agreement shall be invalid, the validity of the remaining provisions of this Agreement shall not be affected. This Agreement is the entire and exclusive agreement between Addinsoft and you with respect to the Software and supersedes all prior agreements (whether written or oral) and other communications between Addinsoft and you with respect to the Software.

Paris, FRANCE, March 2018

System configuration

XLSTAT runs under the following operating systems: Windows Vista, Windows 7, Windows 8.x and 10, Mac OSX 10.6 till 10.12. 32 and 64 bits platforms are supported.

To be able to run XLSTAT required that Microsoft Excel is also installed on your computer. XLSTAT is compatible with the following Excel versions on the Windows systems: Excel 97 (8.0), Excel 2000 (9.0), Excel XP (10.0), Excel 2003 (11.0), Excel 2007 (12.0), Excel 2010 (14.0), Excel 2013 (15.0) and Excel 2016 (16.0) (32 and 64 bits). On the Mac OSX system, XLSTAT is compatible with Excel versions 2011 (14.1 and later) and 2016 (15.27 and later).

Free patches and upgrades for Microsoft Office are available for free on the Microsoft Website. We highly recommend that you download and install these patches as some of them are critical. To check if your Excel version is up to date, please go from time to time to the following web site:

https://docs.microsoft.com/en-us/officeupdates/

Installation

To install XLSTAT you need to:

- Either double-click on the xlstat.exe (PC) or xlstat.dmg (Mac) file that you downloaded from the XLSTAT website www.xlstat.com or from one of our numerous partners.

If your rights on your computer are restricted, you should ask someone that has administrator rights on the machine to install the software for you. Once the installation is over, the administrator must let you have read and write access to the following folder:

- The folder where the XLSTAT user files are located (typically C:\Documents and settings\User Name\Application Data\Addinsoft\XLSTAT\), including the corresponding subfolders.

This folder can be changed by the administrator, using the options dialog box of XLSTAT.

Advanced installation

XLSTAT is easy to deploy within organizations thanks to a variety of functionalities that assist you during the installation on a server, a farm of computers or on computers with multiple user accounts.

In this section:

Silent installation by InstallShield Script (Windows only)

XLSTAT uses an installation program that was created with InstallShield. It is based on install script only. That means that, as with any other installation package based on InstallShield, you do a silent installation.

During the installation, XLSTAT needs that MS Excel is installed on the computer. Excel will be called once to add the XLSTAT button in the Excel main icon bar. The reverse operation is performed during the uninstall process.

Use of an InstallShield script:

You can call the installation program to run a silent installation with the following options that are described in the help of InstallShield.

/uninst: This option forces an uninstall of XLSTAT.

/s: The installation will be done without showing the user dialogs.

/f1 "script file": This parameter indicates the script file that should be used with an absolute path and file name.

/f2 "log file": This parameter indicates the log file that should be used with an absolute path and file name.

/r: This parameter activates the record mode to create a script file.

/L: This parameter allows the selection of the language used during the installation. 10 languages are currently supported as indicated in the following table:

Option	Language
/L1033	English
/L1036	French
/L1031	German
/L1040	Italian
/L1034	Spanish
/L2070	Portugese
/L1045	Polish
/L1041	Japanese
/L1028	traditional Chinese
/L2052	simplified Chinese

/servername=XLSTATLICENSESERVER: this parameter gives the name of the network on which the XLSTAT server is hosted. It is only useful in the case of an XLSTAT client server concurrent license. In that case, XLSTATLICENSESERVER should be replaced by the host name of the server where the XLSTAT concurrent license is hosted.

After the installation of XLSTAT there are two sample script files for installation and uninstall of XLSTAT in the folder silentinstall under the XLSTAT installation folder. You need also the file setup.exe of the installation package to be able to work with the scripts. You obtain these scripts by dezipping the xlstat.zip file that you can download on our website.

To work in a convenient way with scripts for a silent installation, in the following examples, we suppose that the script files and the setup.exe file are located in the same MYDir folder, which is at the same time the current working folder.

Silent installation of XLSTAT

A call to install XLSTAT can be as follows:

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

In this case the script file setup.iss contains the following text:

[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:\My 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=19.1.08.2810

Company=Addinsoft

Lang=040c

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

Result=1

bOpt1=0

bOpt2=0

In this example you may replace the path "C:\Program Files\Addinsoft\XLSTAT" by your desired installation path. You can as well change the path for the user's files "C:\Program Files\Addinsoft\" to a path of your choice.

Silent uninstall of XLSTAT

A call to uninstall XLSTAT can be as follows:

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

In this case the script file setupRemove.iss contains the following text:

[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

[Application]

Name=XLSTAT 2017

Version=19.1.0001

Company=Addinsoft

Lang=0009

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

Result=1

bOpt1=0

bOpt2=0

Silent install of XLSTAT server when using a network concurrent license

Silent installation of XLSTAT Server.

A call to install XLSTAT can be as follows:

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

In this case the script file setup.iss contains the following text:

[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:\My 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=19.1.08.2810

Company=Addinsoft

Lang=040c

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

Result=1

bOpt1=0

bOpt2=0

In this example you may replace the path "C:\Program Files\Addinsoft\XLSTAT" by your desired installation path. You can as well change the path for the user's files "C:\My Documents\Addinsoft\" to a path of your choice.

Silent install of XLSTAT Client on the user computer when using a network concurrent license

A call to install XLSTAT can be as follows:

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

In this case the script file setup.iss contains the following text:

[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:\My 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=19.1.08.2810

Company=Addinsoft

Lang=040c

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

Result=1

bOpt1=0

bOpt2=0

In this example you may replace the path "C:\Program Files\Addinsoft\XLSTAT" by your desired installation path. You can as well change the path for the user's files "C:\My Documents\Addinsoft\" to a path of your choice.

You must enter the hostname of the server where the XLSTAT server license is installed, by replacing "XLSTATLICENSESERVER"by that hostname.

Creating a user defined script file

For further changes to the installation you may also record a manual installation of XLSTAT to create a script file that will be used later. Please use the option /r. A sample call for a script creation might look as follows:

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

Language selection

In most cases, a language selection is not necessary during a silent installation. If XLSTAT was already installed on the computer, the language selection by the installation option /L or the registry entry explained hereunder will have no effect. Each user of the computer will find the language choice he has made before. The user might change the language option at any moment using the XLSTAT Options menu. A demonstration on how a user can change the language is available at:

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

If XLSTAT is being installed for the first time by a user, with the InstallShield interface, then the language that has just been selected for the installation, will be chosen as the default language for XLSTAT.

If XLSTAT is being installed for the first time using a silent installation, then English will be selected as default language. There are two possibilities to change the interface language of XLSTAT before the first start of XLSTAT.

- **/L**: Use this option when calling the silent installation to set the desired language for the installation and for XLSTAT.
- **Register entry**: After the installation of XLSTAT has finished and before XLSTAT is started for the first time, you may change the value of the registry key HKEY_LOCAL_MACHINE\SOFTWARE\XLSTAT+\General\Language to one of the 7 values to set the language of XLSTAT:

Code	Language
US	English
FR	French
DE	German
IT	Italian
ES	Spanish
PT	Portugese
PL	Polish
JP	Japanese
CN	traditional Chinese
CS	simplified Chinese

Selection of the user folder

XLSTAT gives the user the possibility to save data selections and choices made in the dialog boxes that correspond to the different functions, so that you can reuse them during a future session. Further details on how to control this feature can be found in the XLSTAT Options dialog box.

Standard installation of XLSTAT

The selection of the user folder during a standard installation of XLSTAT is set by InstallShield to:

%USERPROFILE%\Application data\ADDINSOFT\XLSTAT

%USERPROFILE%, which is a Windows environment variable, is replaced by its current value during the installation.

Each user has the possibility to change this default value to a user defined value using the corresponding option in the "Advanced" tab of the XLSTAT Options dialog box.

Furthermore you have the possibility to directly change the value of the following registry entry to the desired user folder. The registry entry has priority over the selection in the XLSTAT Options dialog box. The registry entry is different for each user. It has the following name:

HKEY_CURRENT_USER\Software\XLSTAT+\DATA\UserPath

The value of the registry entry may contain environment variables.

Multi-user environment

There are different types of multi-user environments. One example would be a server installation in the case of the Windows Terminal Server or in the case of a Citrix Metaframe Server. Another type of environment is a pool of computers that have all the same installation, often created using an image that has been replicated on all the computers of the pool where some users are authorized to work with XLSTAT. For such cases, please take note of the following advices regarding the choice of the user directories.

In that case, for each user, the user folder should point to a personal folder, for which the user has read and write rights.

There are basically two ways to meet these requirements:

- Use of a virtual folder;
- Use of environment variables.

Virtual folder

In this case, a virtual user folder already exists and is being used. This folder has the same name for every user, but it points to a different folder. A virtual folder is often associated to a user disc like U or X. During the login this user drive is often mounted automatically by a script. The users have normally read and write rights in this folder. For XLSTAT are no further actions necessary regarding the access rights.

If for instance the virtual user folder is **U**, then you can choose the following XLSTAT user folder that will contain the user data following the Microsoft naming conventions:

U:\Application Data\ADDINSOFT\XLSTAT

This folder should exist for each possible XLSTAT user before starting XLSTAT. If this is not the case, an error message informs about the non existing user folder and invites the user to select another user folder.

Environment variables

With this method the value of an environment variable is used to choose a different folder for each user. The user must have read and write rights in that folder.

For instance the environment variable **%USERPROFILE%** can be used to define the following folder using the Microsoft naming conventions:

%USERPROFILE%\Application Data\ADDINSOFT\XLSTAT

The use of environment variables in the dialog boxes of InstallShield is not possible. You may use environment variables in a script file or directly in registry entries.

Server installation and image creation

Server installation and image creation should be possible without any problem. Please notice that Microsoft Excel must have been installed on the machine including all options for VBA (Visual Basic for Applications), Microsoft Forms and graphical filters. During a server installation under Windows Terminal Server, Microsoft Excel version 2003 or later is a preferable choice.

During the installation of XLSTAT, read and write rights are necessary for the folder where the Excel.exe file is located.

If you have specific questions regarding the server installation, do not hesitate to contact the XLSTAT Support.

References

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

Using XLSTAT

The XLSTAT approach

The XLSTAT interface totally relies on Microsoft Excel, whether for inputting the data or for displaying the results. The computations, however, are completely independent of Excel and the corresponding programs have been developed with the C++ programming language.

In order to guarantee accurate results, the XLSTAT software has been intensively tested and it has been validated by specialists of the statistical methods of interest.

Addinsoft has always been concerned about permanently improving the XLSTAT software suite, and welcomes any remarks and improvements you might want to suggest. To contact Addinsoft, write to support@xlstat.com.

Data selection

As with all XLSTAT modules, the selecting of data needs to be done directly on an Excel sheet, preferably with the mouse. Statistical programs usually require that you first build a list of variables, then define their type, and at last select the variables of interest for the method you want to apply to them. The XLSTAT approach is completely different as you only need to select the data directly on one or more Excel sheets.

Three selection modes are available:

- **Selection by range**: you select with the mouse on the Excel sheet all the cells of the table that corresponds to the selection field of the dialog box.
- **Selection by columns**: this mode is faster but requires that your data set starts on the first row of the Excel sheet. If this requirement is fulfilled you may select data by clicking on the name (A, B, …) of the first column of your data set on the Excel sheet, and then by selecting the next columns by leaving the mouse button pressed and dragging the mouse cursor over the columns to select.
- **Selection by rows**: this mode is the reciprocal of the "selection by rows" model. It requires that your data set starts on the first column (A) of the Excel sheet. If this requirement is fulfilled you may select data by clicking on the name (1, 2, …) of the first row of your data set on the Excel sheet, and then by selecting the next rows by leaving the mouse button pressed and dragging the mouse cursor over the rows to select.

Notes:

- Doing multiple selections is possible: if your variables go from column B to column G, and if you do not want to include column E in the selection, you should first select columns B to D with the mouse, then press the Ctrl key, and then select columns F to G still pressing Ctrl. You may also select columns B to G, then press Ctrl, then select column E.
- Multiple selections with selection by rows cannot be used if the transposition option is not activated ( button).
- Multiple selections with selection by columns cannot be used if the transposition is activated ( button).
- When selecting a variable or a group of variables (for example the quantitative explanatory variables) you cannot mix the selection mode. However you may use different modes for different selections within a dialog box.
- If you selected the name of the variables within the data selection, you should make sure the "Columns labels" or "Labels included" option activated.
- You can use keyboard shortcuts to quickly select data. Notice this is possible only you installed the latest patches for Microsoft Excel. Here is a list of the most useful selection shortcuts:
- **Ctrl A**: Selects the whole spreadsheet

- **Ctrl Space**: Selects the whole column corresponding to the already selected cells
- **Shift Space**: Selects the whole row corresponding to the already selected cells
- When one or more cells are selected:
- **Shift Down**: Selects the currently selected cells and the cells on the row below on one row
- **Shift Up**: Selects the currently selected and the cells on the row below on one row
- **Shift Left**: Selects the currently selected and the cells to the left on one column
- **Shift Right**: Selects the currently selected and the cells to the right on one column
- **Ctrl Shift Down**: Selects all the adjacent non empty cells below the currently selected cells
- **Ctrl Shift Up**: Selects all the adjacent non empty cells above the currently selected cells
- **Ctrl Shift Left**: Selects all the adjacent non empty cells to the left of the currently selected cells
- **Ctrl Shift Right**: Selects all the adjacent non empty cells to the right of the currently selected cells
- When one ore more columns are selected:
- **Shift Left**: Selects one more column to the left of the currently selected columns
- **Shift Right**: Selects one more column to the right of the currently selected columns
- **Ctrl Shift Left**: Selects all the adjacent non empty columns to the left of the currently selected columns
- **Ctrl Shift Right**: Selects all the adjacent non empty columns to the right of the currently selected columns
- When one or more rows are selected:
- **Shift Down**: Selects one more row to the left of the currently selected rows
- **Shift Up**: Selects one more row to the right of the currently selected rows
- **Ctrl Shift Down**: Selects all the adjacent non empty rows below the currently selected rows
- **Ctrl Shift Up**: Selects all the adjacent non empty rows above the currently selected rows

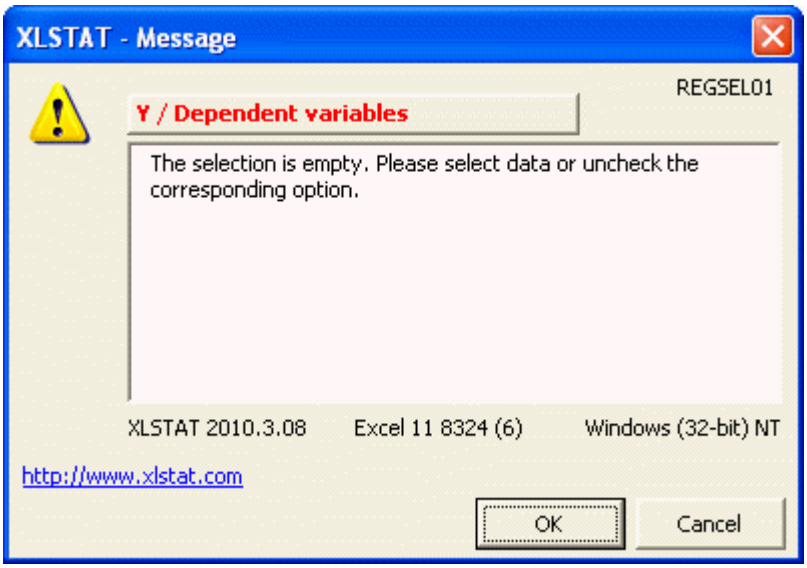
See also:

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

Messages

XLSTAT uses an innovative message system to give information to the user and to report problems.

The dialog box below is an example of what happens when an active selection field (here the Dependent variables) has been activated but left empty. The software detects the problem and displays the message box.



The information displayed in red (or in blue depending on the severity) indicates which object/option/selection is responsible for the message. If you click on OK, the dialog box of the method that had just been activated is displayed again and the field corresponding to the Quantitative variable(s) is activated.

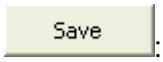
This message should be explicit enough to help you solve the problem by yourself. If a tutorial is available, the hyperlink "http://www.xlstat.com" links to a tutorial on the subject related to the problem. Sometimes an email address is displayed below the hyperlink to allow you send an email to Addinsoft using your usual email software, with the content of the XLSTAT message being automatically displayed in the email message.

General options

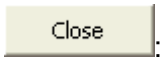
Options

XLSTAT offers several options in order to allow you to customize and optimize the use of the software.

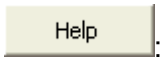
To display the options dialog box of XLSTAT, click on "Options" in the menu or on the  button of the XLSTAT toolbar.



: Click this button to save the changes you have made.



: Click this button to close the dialog box. If you haven't previously saved the options, the changes you have made will not be kept.



: Click this button to display the help.


: Click this button to reload the default options.

General tab:

Language: Use this option to change the language of the interface of XLSTAT.

Focus: This option is only visible if you are using the trial version or the premium version of XLSTAT. In this case, you can use this option to display a button giving a quicker access to a series of functions corresponding to one of the fields covered by the "Applied" solutions of XLSTAT: Forecasting, LifeScience, Marketing, Quality, Sensory.

Dialog box entries:

- **Memorize during one session**: Activate this option if you want XLSTAT to memorize during one cession (from opening until closing of XLSTAT) the entries and options of the dialog boxes.
- **Including data selections**: Activate this option so that XLSTAT records the data selections during one session.
- **Memorize from one session to the next**: Activate this option if you want XLSTAT to memorize the entries and options of the dialog boxes from one session to the next.
- **Including data selections**: Activate this option so that XLSTAT records the data selections from one session to the next. This option is useful and saves your time if you work on spreadsheets that always have the same layout.

Ask for selections confirmation: Activate this option so that XLSTAT prompts you to confirm the data selections once you clicked on the OK button. If you activate this option, you will be able to verify the number of rows and columns of all the active selections.

Show only the active functions in menus and toolbars: Activate this option if you want that only the active functions corresponding to registered modules are displayed in the XLSTAT menu and in the toolbars.

Show the 'Start XLSTAT' button in the Excel ribbon: Activate this option so that when Excel starts, the 'Start XLSTAT' button is displayed in the Excel ribbon. Clicking the 'Start XLSTAT' button will load XLSTAT and the full XLSTAT ribbon.

Start XLSTAT with Excel: Activate this option so that XLSTAT is launched when Excel starts.

Display the tutorials during the XLSTAT startup: Activate this option so that step by step tutorials are suggested when XLSTAT starts.

Delete hidden sheets when they are no longer used: Some XLSTAT functions require hidden sheets to store data that are used to create charts. Activate this option, so that when you delete a sheet where there are some XLSTAT results, the corresponding hidden sheets are deleted as well.

Data tab:

Excel filters: Use the following options to set how XLSTAT should handle data filters, that have been applied to your worksheet, when they exist.

- **Ask the user**: Activate this option if you want XLSTAT to prompt you whenever a filter is found. You will be offered the choice between using the filters as they are or ignoring them.
- **Apply filters as they are on the worksheet**: Activate this option so that XLSTAT takes into account the filters that have been applied on the worksheet.
- **Ignore filters**: Activate this option if you want XLSTAT to ignore filters and use the whole dataset.

Missing data:

Consider empty cells as missing data: this is the default option for XLSTAT and it cannot be changed. Empty cells are considered by all tools as missing data.

Consider also the following values as missing data: when a cell contains a value that is in the list, below this option, it will be considered as a missing data, whether the corresponding selection is for numerical or categorical data.

Consider all text values as missing data: when this option is activated, any text value found in a table that should contain only numerical values, will be converted and considered by XLSTAT as a missing data. This option should be activated if you are sure that text values can not correspond to numerical values converted to text by mistake.

Outputs tab:

Position of new sheets: If you choose the "Sheet" option in the dialog boxes of the XLSTAT functions, use this option to modify the position if the results sheets in the Excel workbook.

Color tabs: Activate this option if you want to highlight the tabs produced by XLSTAT using a specific color.

Display the report header:

- **Display the results list in the report header**: Activate this option so that XLSTAT displays the results list at the bottom of the report header.
- **Display the project name in the report header**: Activate this option to display the name of your project in the report header. Then enter the name of your project in the corresponding field.
- **Display the action buttons**: Activate this option to display the buttons that allow to relaunch an analysis.

Display comments: Activate this option if you want that XLSTAT displays comments that help you with interpreting the results, when available. Comments can be seen by hovering the mouse cursor over the small red triangles displayed to the left of the titles of each

Merge cells to the left of results tables: Activate this option to merge the cells that to the left of results tables. This will allow you to easily add your own interpretation of the results.

Display an outline: Activate this option to use the Excel feature that allows creating outlines in a spreadsheet. Outlines can make it easier for to consult the report.

- **Collapsed by default**: Activate this option so that the outlines of the report are all collapsed when the report is displayed.

Enlarge the first column of the report by a factor of X: Enter the value of the factor that is used to automatically enlarge the width of the first column of the XLSTAT report. Default value is 1. When the factor is 1 the width is left unchanged.

Display titles in bold: Activate this option so that XLSTAT displays the titles of the results tables in bold.

Empty rows after titles: Choose the number of empty rows that must be inserted after titles. The number of empty rows after tables and charts corresponds to this same number +1.

Theme for tables: Choose the theme to apply on the tables produced by XLSTAT.

Display table headers in bold: Activate this option to display the headers of the results tables in bold.

Number of decimals: Choose the number of decimals to display for the numerical results. Notice that you always have the possibility to view a different number of decimals afterwards, by using the Excel formatting options.

Minimum p-value: Enter the minimum p-value below which the p-values are replaced by "< p" where p is the minimum p-value.

Charts tab:

Style: Choose the output style that fits your needs. "Classic" corresponds to the style XLSTAT has been using since its inception. "Modern" uses a different palette of colors. Scientific is using black, white and grey colors.

Display charts on separate sheets: Activate this option if you want that the charts are displayed on separate chart sheets. Note: when the charts are displayed on a spreadsheet you can still transform them into a chart sheet, by clicking the right button of the mouse, and then selecting "location" and then "As new sheet".

Charts size:

- **Automatic**: Choose this option if you want XLSTAT to automatically determine the size of the charts using as a starting value the width and height defined below.
- **User defined**: Activate this option if you want XLSTAT to display charts with dimensions as defined by the following values:
- **Width**: Enter the value in points of the chart's width;
- **Height**: Enter the value in points of the chart's height.

Display charts with aspect ratio equal to one: Activate this option to ensure that there is no distortion of distances due to different scales of the horizontal and vertical axes that could lead to misinterpretations.

Advanced tab:

Random numbers:

Fix the seed to: Activate this option if want to make sure that the computations involving random numbers always give the same result. Then enter the seed value.

Maximum nu mber of processors: XLSTAT can run calculations on multiple processors to educe the computing time. Choose the maximum number of processors that XLSTAT can use.

Show the advanced buttons in the dialog boxes: Activate this option if you want to display the buttons that allow to save or load dialog box settings, or generate VBA code to automate XLSTAT runs.

Participate in the continuous improvement of XLSTAT: Activate this option if you agree to participate in the continuous improvement of XLSTAT by reporting crashes if they happen and the name of the methods you use. Data are never transmitted.

Path for the user's files: This path can be modified if and only if you have administrator rights on the machine. You can then modify the folder where the user's files are saved by clicking the […] button that will display a box where you can select the appropriate folder. User's files include the general options as well as the options and selections of the dialog boxes of the various XLSTAT functions. The folder where the user's files are stored must be accessible for reading and writing to all types of users.

XLSTAT-R tab:

Hide XLSTAT-R: If you check this option, the XLSTAT-R menu will not be displayed in the ribbon.

RScript.exe: So that XLSTAT-R can run, the location of RScript.exe must be known by XLSTAT. Should it not have been automatically detected by XLSTAT, you must specify where the file is located.

XML editor: XLSTAT-R uses XML files to exchange information with the R programs. You can create your own connectors. You can specificy in this field the program you want to use to create or edit the XML files.

CRAN Mirror: Use this option to let XLSTAT know which CRAN Mirror you want to use to download the latest version of the R packages. Default value is https://cran.revolutionanalytics.com.

Groups and R packages: Check in the list the groups and R packages you want to use through XLSTAT.

Notifications tab:

Notify me before the license or the access to upgrades expires: Activate this option should you want XLSTAT to warn you before your license expires.

Display information related to: Select the subjects that you want XLSTAT to inform you about.

Workflow

The workflow feature allows you to combine and run multiple analyses in a row. Each one can use the results of previous analyses as input data as well as data from open Excel workbooks. The sequence of data and statistical analyses is thus smoothed and presented in a simpler way. This allows you to have an overall view of the analyses while keeping the possibility to replay each analysis independently. Export/import features of the channels allow to share easily the configurations

In this section :

Description

A workflow is a succession of analyses represented by interconnected blocks. Visually, it looks like a tree that evolves from left to right. The first blocks (on the left) correspond to the input data and all the following blocks are the analyses that are chained one after the other, in the order represented by the tree. The last blocks are the final result of the chain. For each analysis block, the input data can be selected from the output data of all the previous blocks. It is sufficient that they belong to the same branch. If data is filtered then this does not prevent it from being selected. Depending on your settings in the XLSTAT options, the filters will be applied or not.

Input data: 

- You can select these, via the input data block, from all the data ranges of the open Excel workbooks. This selection can be manual or automatic. A tool allows you to automatically detect data from open workbooks and reuse them for future analyses.
- No block can precede an input data block.
- If data is filtered then this does not prevent it from being selected. Depending on your settings in the XLSTAT options, the filters will be applied or not.
- You can choose to have a fixed number of rows or not. This allows you to stay on a specific selection of data or to let the tool add/remove rows according to your actions on the input data.
- The same thing is possible for the columns. However, if you delete columns that are used in the rest of your workflow, then the blocks concerned and those affected will be reinitialized with the deletion of the result sheet if it exists.

See section Create a new workflow for more details on these points.

Available analyses

Twenty-five analyses are currently available to create a pathway. All of them can be used in succession. However, Descriptive statistics, Histograms and Scatterplots cannot be continued, they are necessarily final blocks.

- **Preparing data**:
  -  : Missing data
  -  : Data management
  -  : Data anonymization
  -  : Variables transformation

-  : Create a contingency table

- **Describing data**:
  -  : Descriptive statistics
  -  : Normality test
  -  : Filtered table

- **Visualizing data**:
  -  : Histograms
  -  : Scatter plots

- **Analyzing data**:
  -  : Principal Component Analysis (PCA)
  -  : k-means clustering
  -  : Fuzzy k-means clustering
  -  : Agglomerative hierarchical clustering (AHC)

- **Modeling data**:
  -  : Linear regression
  -  : ANOVA
  -  : Logistic regression
  -  : PLS Regression

- **Time series analysis**:
  -  : ARIMA
  -  : Time series transformation
  -  : Mann-Kendall trend tests

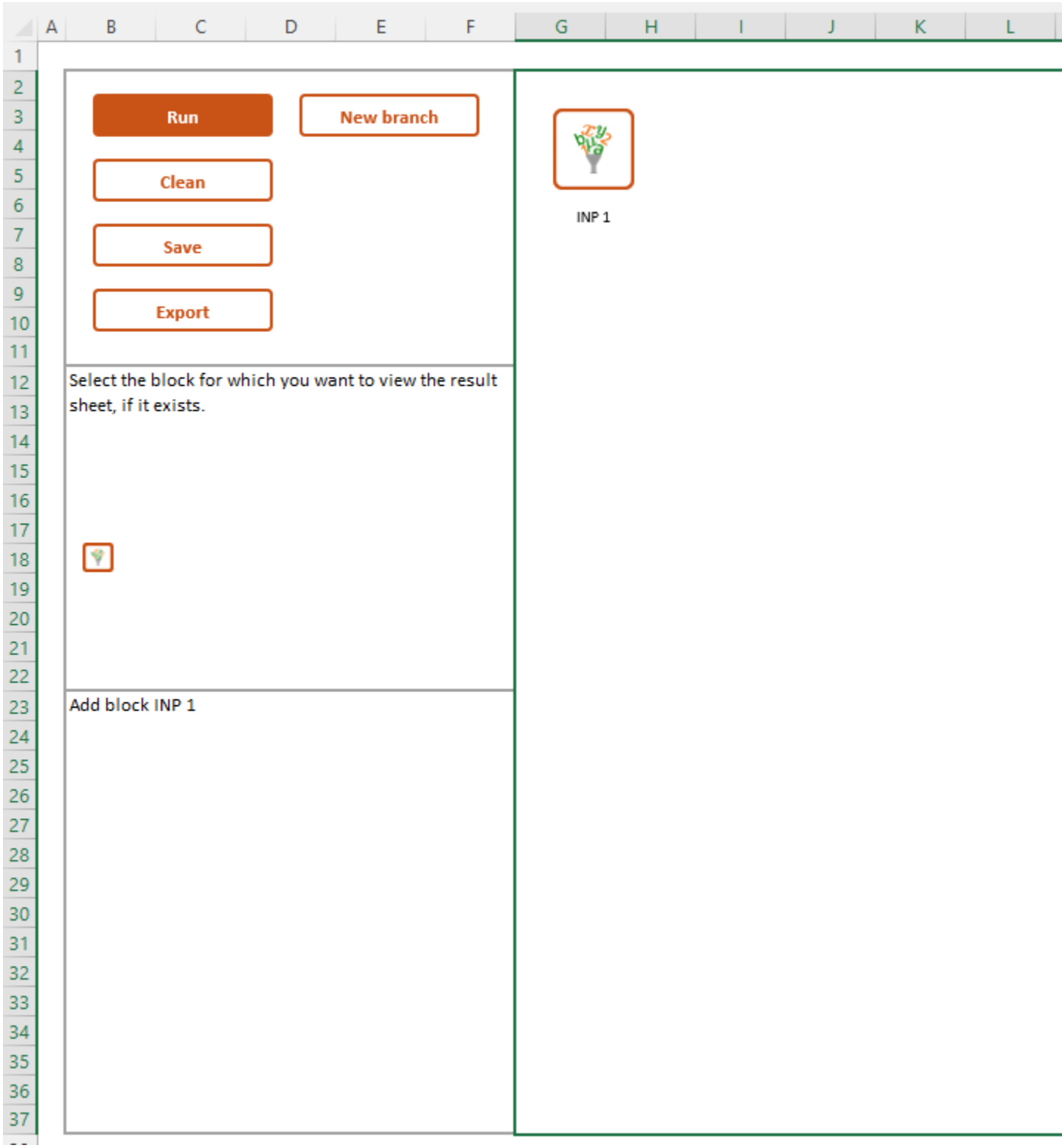
- **Sensory data analysis**:
  -  : Variance by class"
  -  : CATA data analysis

- **Text mining**:
  -  : Cleaning text data
  -  : Feature extraction

This list is subject to change.

Workflow screen

The workflow screen looks like the figure below.



Here are the details of the different parts of the screen:

- **Part 1 (top left)**: it contains various buttons of which here are the details:

- **Run**: click on this button if you want to run all the calculations of the workflow. Each block, from the input data to the final analyses, will be launched in order to visualize the final result(s). If an analysis does not run correctly then the calculations stop at that analysis and the user is warned. It is possible to modify the input data external to the workflow (the input data of the blocks related to the input data of the workflow). This change will be taken into account for the new calculations. However, it may happen that this change impacts the number of columns of some output tables. In this case, you will be asked to recalculate the input data of the concerned analyses.

- **Clean**: click on this button if you want to clean up part 4 in order to start from a clean sheet. Please note that all existing blocks (input data and analyses) will be deleted.

- **Save**: click on this button if you want to save your workflow. A .wkf file is created. It is stored in the same location as all your XLSTAT user files (the path can be configured in the XLSTAT advanced options). It contains the structure and parameters of your workflow but not the input data. When you close an Excel workbook containing one or more workflows, you will be asked if you want to save your workflows. Indeed, once the workbook is closed, any unsaved workflow will be lost. If you want to share your workflow with another person then you have to export it by clicking on the **Export** button.

- **Export**: click on this button if you wish to export your workflow. A window will open in order to select the export directory and possibly modify the name of the .wkf file containing all the elements necessary for the reconstruction of the exported workflow. You will be able to share this file with another person who will be able to view and modify the workflow after importing it into an Excel workbook (see Import a workflow). It contains the structure and the parameters of your workflow but also the input data.

- **New branch**: Click on this button if you want to create a new branch. A new block for input data will appear in part 4. It will be on the left and in the first available location because no block can precede it. Its settings window will also appear. You can then follow it with new or existing analyses by connecting it to a branch.

- **Part 2 (middle left)**: it contains a thumbnail of the workflow which is in part 4 and allows you to quickly visualize the input data (if it exists) or the result sheet of an analysis (if it exists). Just click on the thumbnail of the target block.

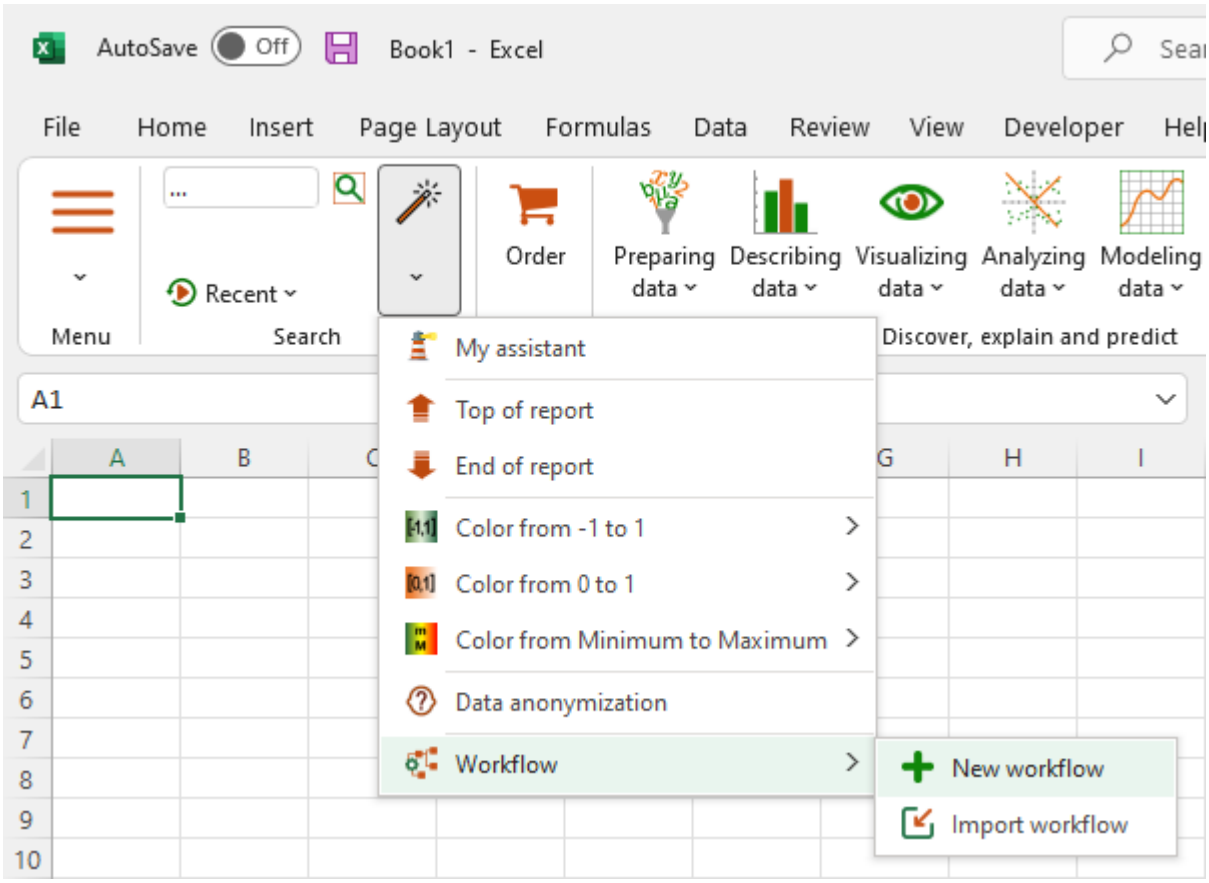
- **Part 3 (bottom left)**: it contains comments that are updated automatically and provide information on the actions performed.

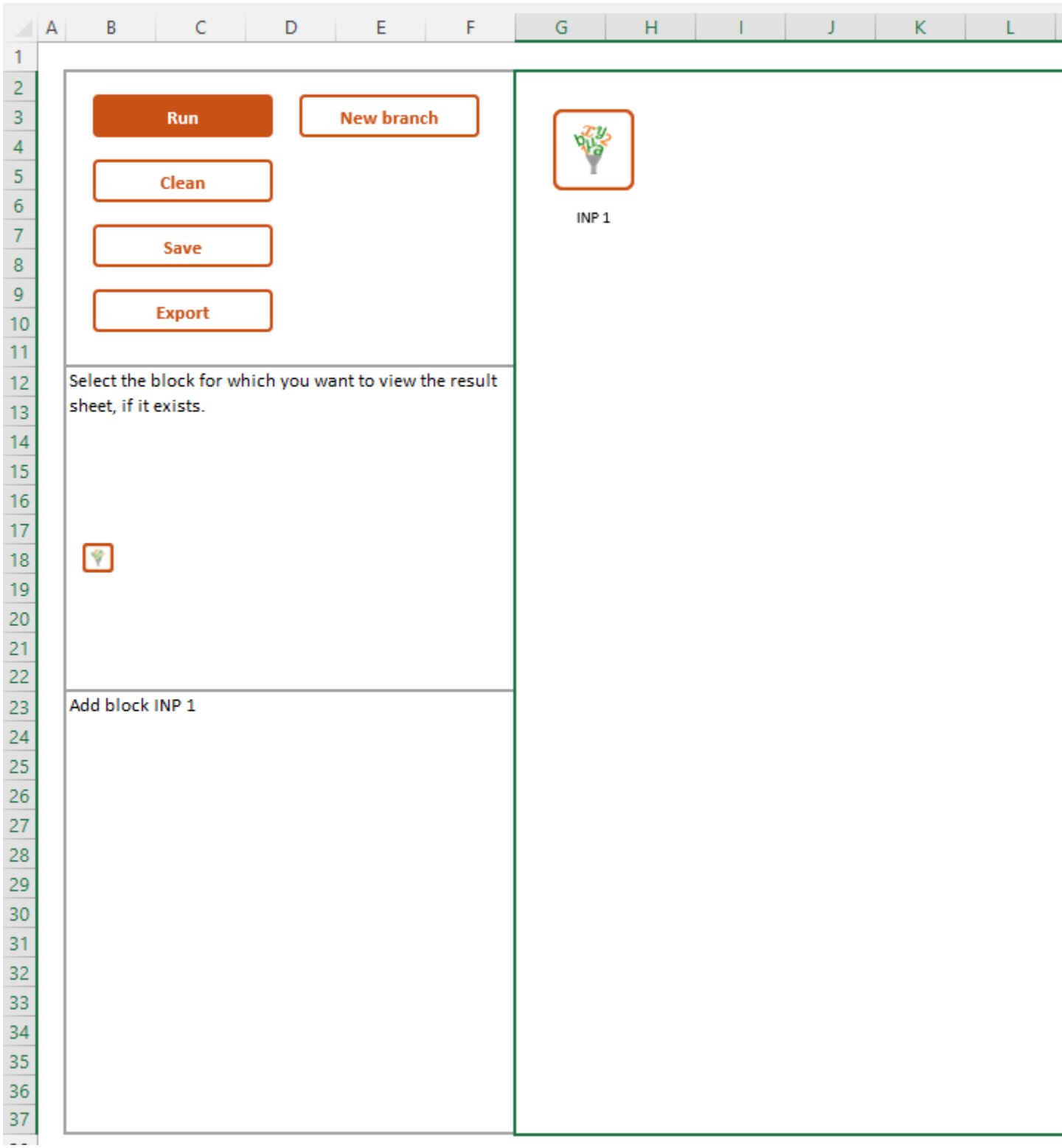
- **Part 4 (right)**: it contains the workflow itself composed of all the action blocks from which some actions can directly be performed.

Create a new workflow

Displaying the workspace of a new workflow

To build a new workflow, you need to start by displaying the associated screen, presented previously in the section Workflow screen. To do this, start XLSTAT and select the workflow tool in the ribbon as shown in the figure below.



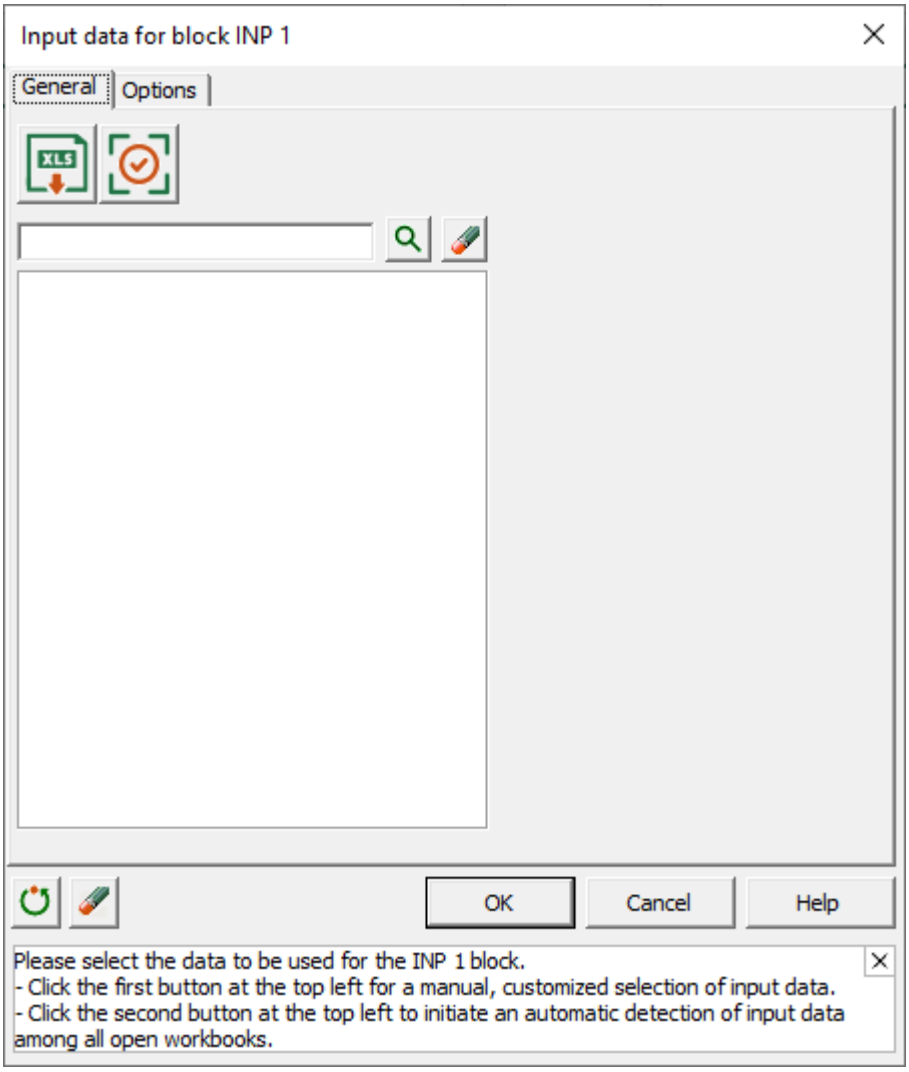


A new Excel sheet will be created with the workspace of the new workflow to be built.

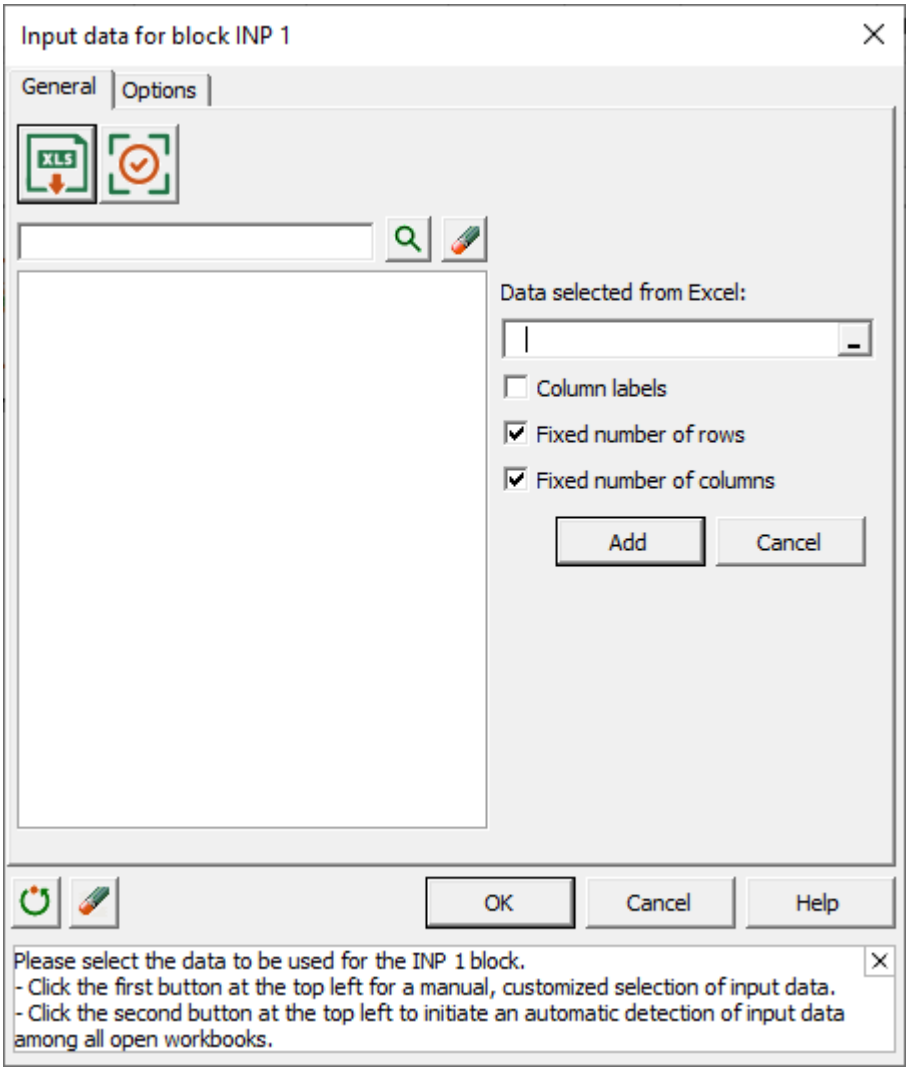
Setting up a first block for input data

We will detail the contents of the setting dialog box for input data.

Onglet **General**:



 : click this button if you want to select the input data manually from all open workbooks.

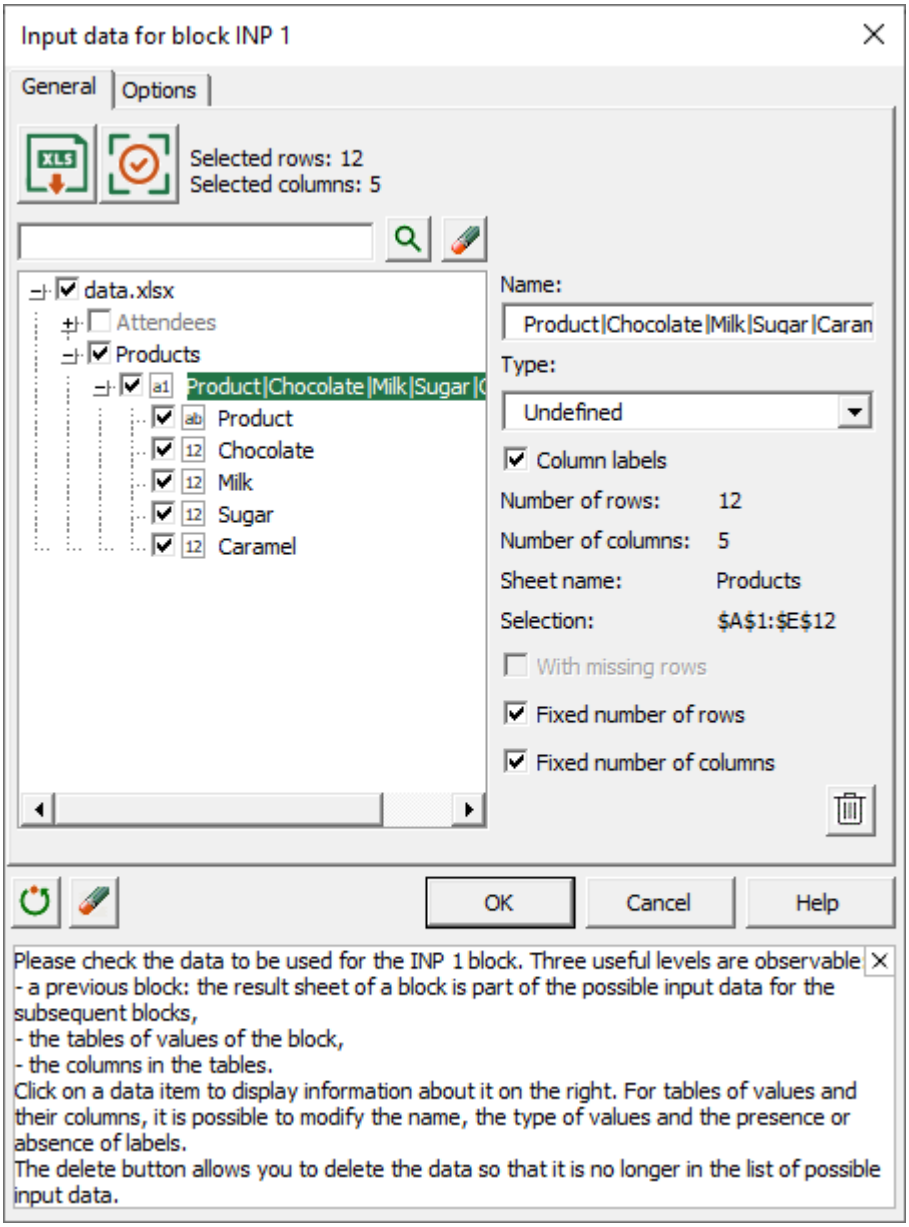


Start by selecting a first table of data in a sheet of an open Excel workbook. You can then update the following information:

- **Column labels**: if checked then it means that the columns have labels.
- **Fixed number of rows**: if checked then this means that new rows added later at the end of the selected Excel table will be automatically added to the block data. Any insertion or deletion of rows within the Excel table will be automatically taken into account, whether the box is checked or not.
- **Fixed number of columns**: if checked then it means that new columns added later at the end of the selected Excel table will be automatically added to the block data. Any insertion or deletion of columns within the Excel table will be automatically taken into account, whether the box is checked or not. Be careful, a deletion of columns used afterwards in the workflow will lead to a reset of the impacted blocks with the deletion of their result sheet if it exists.

Click **Add** to confirm your choice or **Cancel** to leave it blank.

 : click this button if you want to automatically detect possible input data among all open workbooks. This will populate the dialog box with a tree of the different detected ranges.



We can observe four useful levels:

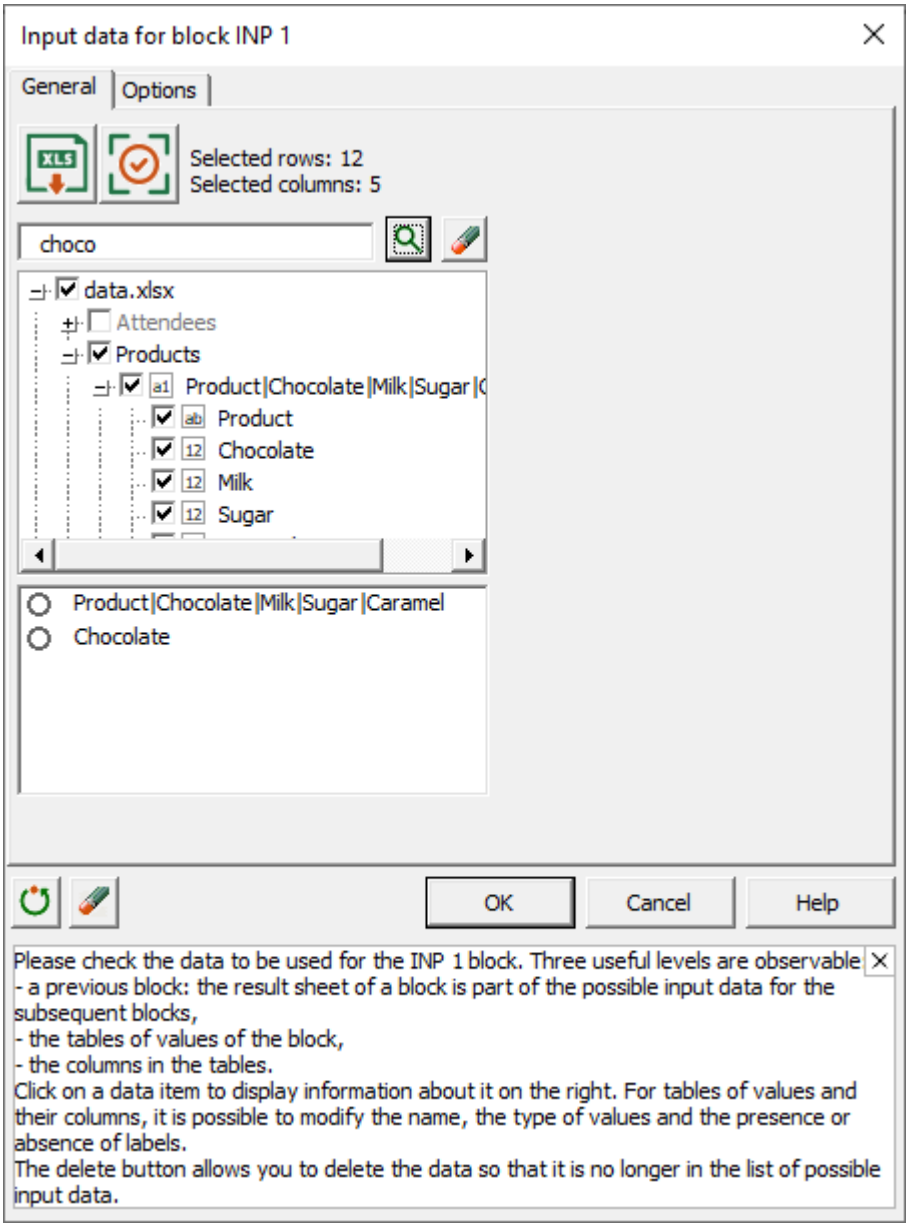
- **Workbook** with the name of the workbook where the data range is located.
- **Worksheet** with the name of the worksheet where the data range is located.
- **Table** with the labels of the columns in the data range (or the first and last cell if no column labels are detected).
- **Column** with the label of the concerned column (or the first and last cell if no column label is detected).

It is possible to click on one of the elements of the tree in order to obtain and possibly modify certain information:

- **Name**: name of the element in the tree structure. It can be modified.
- **Type**: type of the element: undefined, quantitative, qualitative. It can be modified.

- **Column labels**: if checked then it means that the element has a label if it is a column or several if it is a table.
- **Number of rows**
- **Number of columns**
- **Sheet name**
- **Selection**: first and last cells containing the selected data.
- **With missing rows**: if checked then it means that there are empty rows in the element.
- **Fixed number of rows**: if checked then this means that new rows added later at the end of the data range relative to the selected element will be automatically added to it. Any insertion or deletion of rows within the same data range will be automatically taken into account, whether the box is checked or not. This information can only be modified at table level.
- **Fixed number of columns**: if checked then it means that new columns added later at the end of the data range related to the selected item will be automatically added to the block data. Any insertion or deletion of columns within the same data range will be automatically taken into account, whether the box is checked or not. This information can only be modified at table level. Be careful, a deletion of columns used afterwards in the workflow will lead to a reinitialization of the impacted blocks with deletion of their result sheet if it exists.
-  : button to delete the selected item from the tree.

A search bar is located above the tree structure. It allows you to easily find an element within it thanks to the display that appears below the tree structure as in the figure below.



And finally:

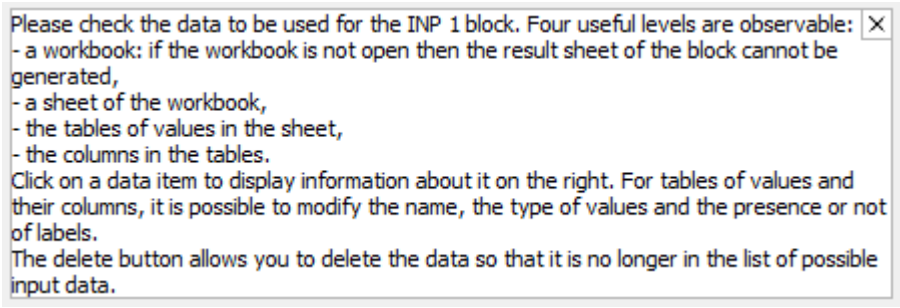
OK: click on this button to validate and save your settings.

Cancel: click on this button to close the dialog box without saving the settings.

Help: Click on this button to display the help related to workflows in XLSTAT.

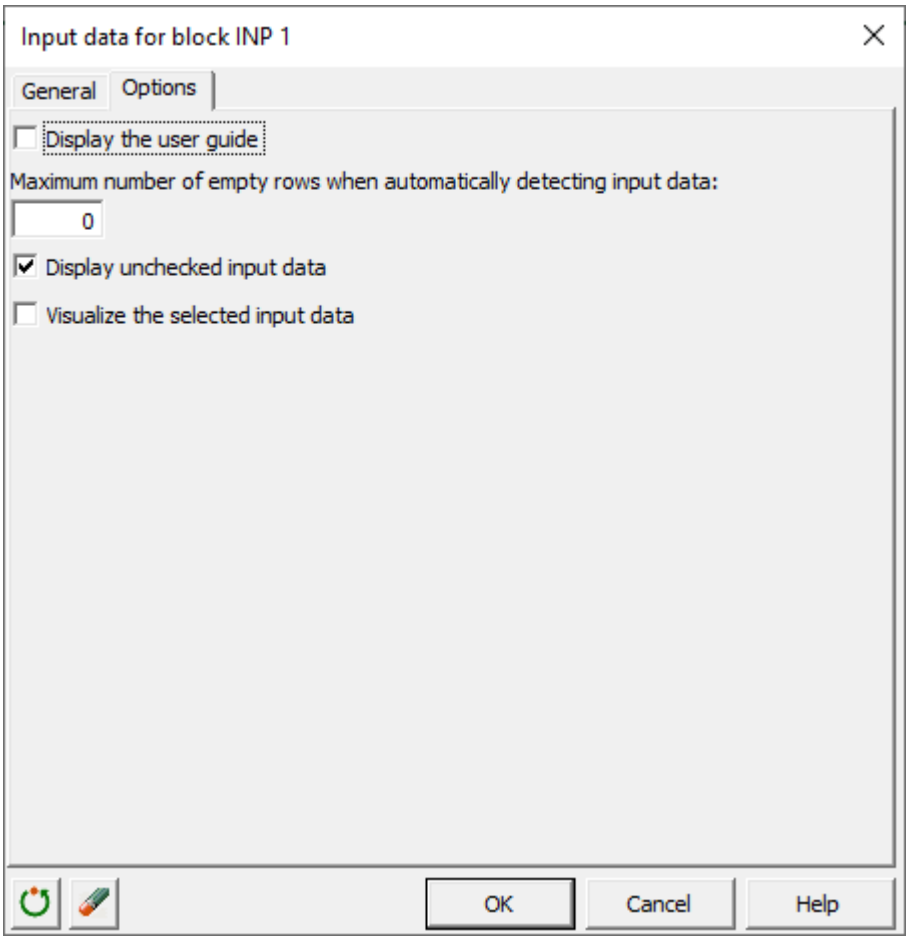
: click on this button to reset the dialog box to the default settings.

: click on this button to clear the data selections from the dialog box.



: This insert serves as a user guide for the actions to be taken. It can be removed by clicking on the cross at the top right of it or via the options.

Onglet **Options**:

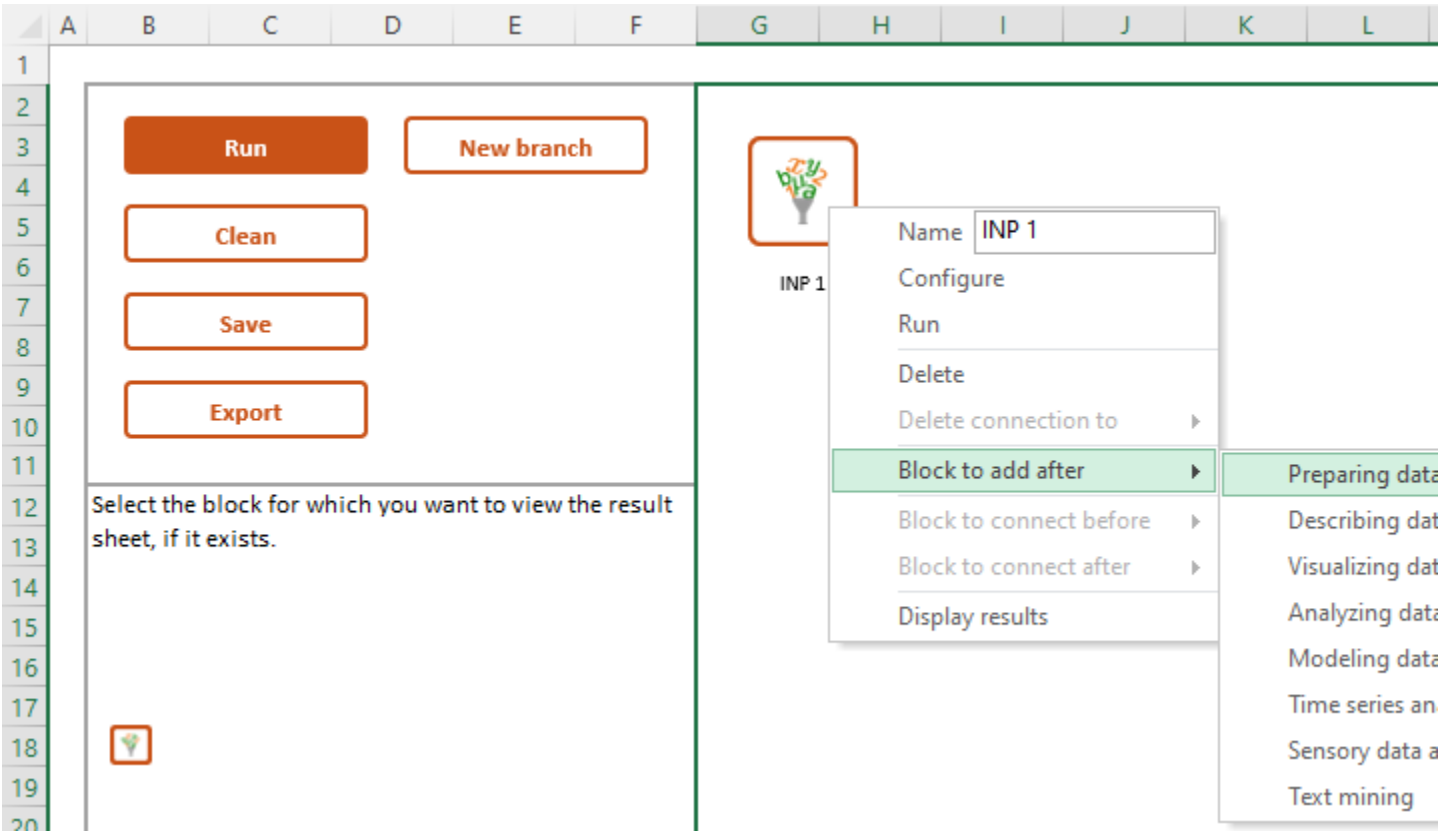


- **Display the user guide**: check this box if you want to display the user guide at the bottom of the dialog box.
- **Maximum number of empty rows when automatically detecting input data**: if you select your input data from the automatic detection tool, you can specify the maximum number of empty lines. If this is 0 (default value) then no empty lines will be accepted.
- **Display unchecked input data**: check this box if you want to display the unchecked input data. In the General tab, especially if you choose to do automatic detection, it is likely that not all data will be of interest to you. Once you have made your selection, you can choose not to display the other data, but still keep them in memory. We have seen above that it is also possible to delete them.

- **Visualize the selected input data**: check this box if you want to view the selected input data in the tree view of the General tab.

Adding a new block

Once the input data block has been added, you can click on it to display the menu of different actions that can be taken from this block. These will be detailed in the section Actions on a workflow block. However, we will focus on the one regarding the addition of a block.

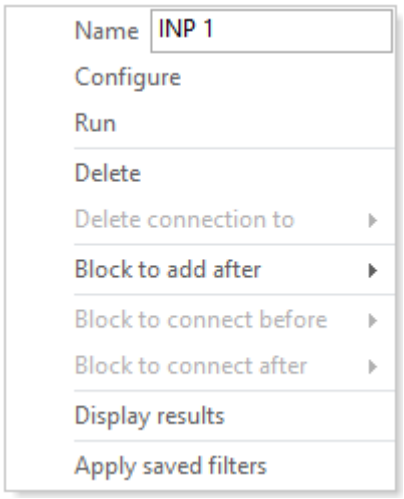


In the menu, go to "Block to add after", then click on the analysis you are interested in. A new block will be added after and the associated dialog box will appear. All you have to do is to set the parameters as you wish. To access the documentation of this one, you can click on the name of the target analysis in the part Different analyses that can be used.

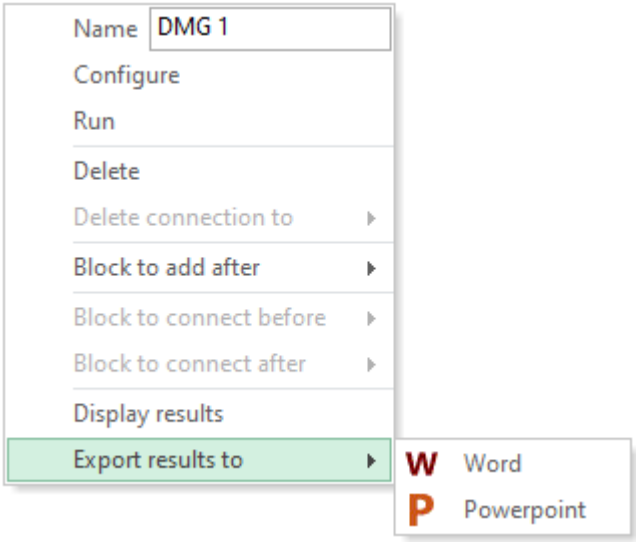
You can add as many blocks as you want. As a reminder, Descriptive statistics, Histograms and Scatter plots cannot have a block after them.

Actions on a workflow block

Menu for an input data block:



Menu for an analysis block:

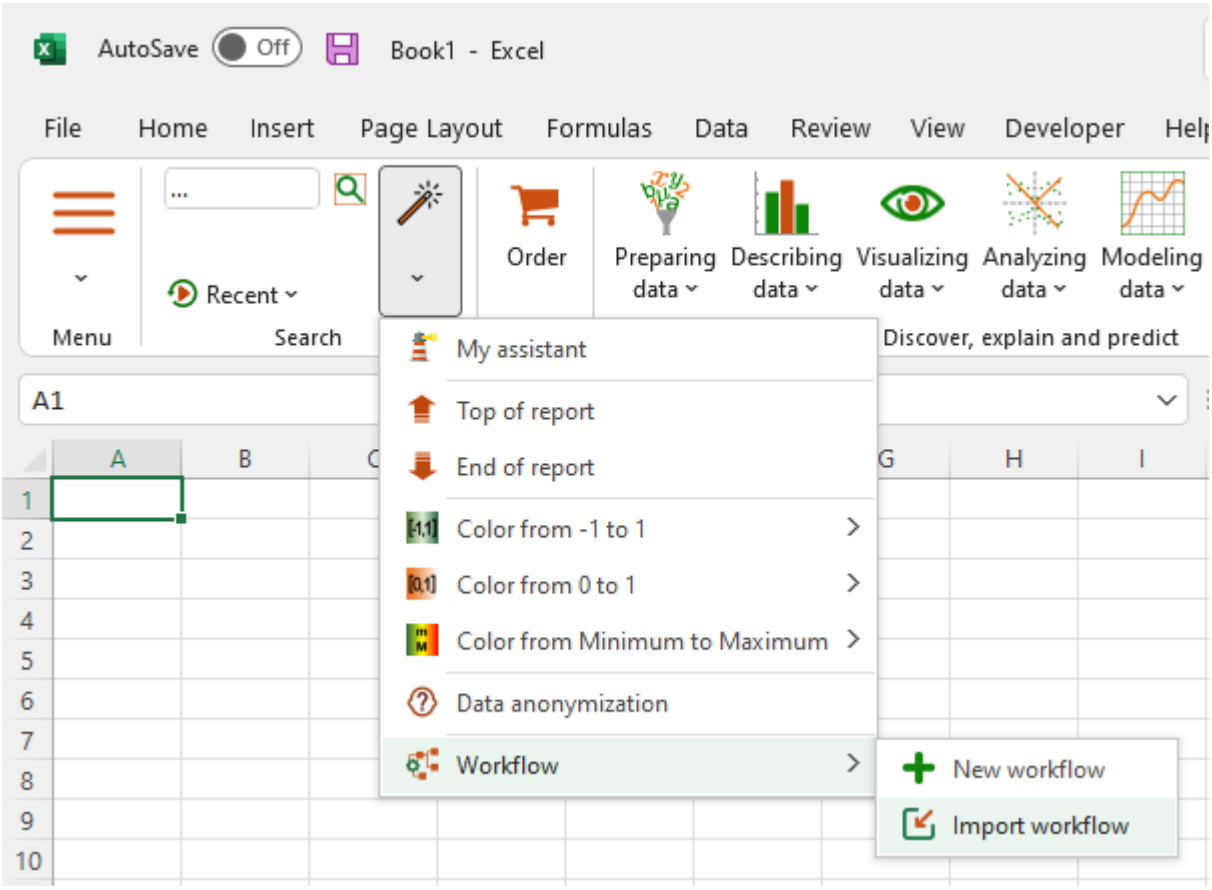


- **Name**: you can change the name of a block via this text box. This will be updated after you click on Enter.
- **Configure**: click on this item if you want to display the dialog box for the block setup. If you configure a block then the blocks connected afterwards can be impacted. In this case the result sheets of the impacted blocks will be deleted and you will be asked to restart the calculations if you want to update them. In a workflow, a green connector means that the block it points to has a result sheet that can be viewed.
- **Run**: click on this item if you want to run the calculations up to the block. All analysis blocks belonging to the same branch, from the leftmost to this one, will have their results updated.
- **Delete**: click on this item if you want to delete the block. All subsequent blocks, with or without a direct connection, will also be deleted.

- **Delete connection to**: select this item and click on the name of the next directly connected block whose you want to remove connection. If this next block has no other previous block then it is deleted and all subsequent blocks, with or without a direct connection, will also be deleted.
- **Block to add after**: select this item and click on the analysis block you want to add next.
- **Block to connect before**: select this item and click on the name of the block you want to connect before. If you can't find the block you want, then it is not one of the possible blocks.
- **Bloc to connect after**: select this item and click on the name of the block you want to connect next. If you can't find the block you want, then it is not one of the possible blocks.
- **Display results**: this item is enabled in the menu only for blocks that have a result sheet. Click on this item if you want to display the result of the analysis associated with the block. You will be sent directly to the result sheet. The calculations will not be restarted.
- **Apply saved filters**: this item is enabled in the menu only for input data blocks that have filtered data. Click on it if you want to apply the filters used for this block in the relevant sheets. It is possible to have several input data blocks with data from the same sheets but with different filters.
- **Export results to**: this item is enabled in the menu only for blocks that have a result sheet. Select it to export the results of the block. Click on Word or Powerpoint and a dialog box will appear to choose the elements of the result sheet to export.

Import a workflow

We have seen that it is possible to export a workflow via the **Export** button on the left in the Workflow tool workspace. This will allow you to import it later. It can also be imported by another person to whom you have sent the .wkf file created during the export. The import button is located in the ribbon at the level of the Workflow tool as shown in the figure below.



Several Excel sheets will be added to the workbook, those related to the dataset used for the imported workflow and another one with the workspace of this one. You will be asked if you want to run the calculations or not.

Examples

Examples of how to use the Workflow tool are available on the XLSTAT Help Center:

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

Preparing data

Data sampling

Use this tool to generate a subsample of observations from a set of univariate or multivariate data.

In this section:

Description

Sampling is one of the fundamental data analysis and statistical techniques. Samples are generated to:

Test an hypothesis on one sample, then test it on another;

Obtain very small tables which have the properties of the original table.

To meet these different situations, several methods have been proposed. XLSTAT offers the following methods for generating a sample of N observations from a table of M rows:

N first rows: The sample obtained is taken from the first N rows of the initial table. This method is only used if it is certain that the values have not been sorted according to a particular criterion which could introduce bias into the analysis;

N last rows: The sample obtained is taken from the last N rows of the initial table. This method is only used if it is certain that the values have not been sorted according to a particular criterion which could introduce bias into the analysis;

N every s starting at k: The sample is built extracting N rows, every s rows, starting at row k;

Random without replacement: Observations are chosen at random and may occur only once in the sample;

Bootstrap (random with replacement): Observations are chosen at random and may occur several times in the sample;

Systematic from random start: From the j'th observation in the initial table, an observation is extracted every k observations to be used in the sample. j is chosen at random from among a

number of possibilities depending on the size of the initial table and the size of the final sample. k is determined such that the observations extracted are as spaced out as possible;

Systematic centered: Observations are chosen systematically in the centers of N sequences of observations of length k;

Random stratified (1): Rows are chosen at random within N sequences of observations of equal length, where N is determined by dividing the number of observations by the requested sample size;

Random stratified (2): Rows are chosen at random within N strata defined by the user. In each stratum, the number of sampled observations is proportional to the relative frequency of the stratum.

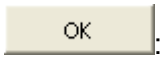
Random stratified (3): Rows are chosen at random within N strata defined by the user. In each stratum, the number of sampled observations is proportional to a relative frequency supplied by the user.

User defined: A variable indicates the frequency of each observation within the output sample.

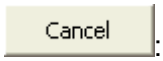
Training and test sets: Data are split into two parts – a training set and a test set. The rows of each set are randomly drawn from the initial dataset. The size of the training set is defined by a number of rows.

Training and test sets (%): Data are split into two parts – a training set and a test set. The rows of each set are randomly drawn from the initial dataset. The size of the training set is defined by a row number percentage from the initial data set.

Dialog box



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

Data: Select the data in the Excel worksheet.

Sampling: Choose the sampling method (see the description section for more details).

Sample size: Enter the size of the sample to be generated.

Strata: This option is only available for the random stratified sampling (2) and (3). Select in that field a column that tell to which stratum each observation belongs.

Weight of each stratum: This option is only available for the random stratified sampling (3). Select a table with two columns, the first containing the strata ID, and the second the weight of the stratum in the final sample. Whatever the weight unit (size, frequency, percentage), XLSTAT standardizes the weight so that the sum is equal to the requested sample size.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (data and observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Display the report header: Deactivate this option if you want the sampled table to start from the first row of the Excel worksheet (situation after output to a worksheet or workbook) and not after the report header. You can thus select the variables of this table by columns.

Shuffle: Activate this option if you want to randomly permute the output data. If this option is not activated, the sampled data respect the order of the input data.

Display side by side: Activate this option to display generated samples side by side.

References

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

Hedayat A.S. & Sinha B.K. (1991). Design and inference in finite population sampling. John Wiley & Sons, New York.

Distribution sampling

Use this tool to generate a data sample from a continuous or discrete theoretical distribution or from an existing sample.

In this section:

Description

Where a sample has been generated from a theoretical distribution, you must choose the distribution and, if necessary any parameters required for this distribution.

Distributions

XLSTAT provides the following distributions:

- Arcsine (a): the density function of this distribution (which is a simplified version of the Beta type I distribution) is given by:

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

We have $E(X) = \alpha$ and $V(X) = \alpha(1-\alpha)/2$

- Bernoulli (p): the density function of this distribution is given by:

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

We have $E(X) = p$ and $V(X) = p(1-p)$

The Bernoulli, named after the Swiss mathematician Jacob Bernoulli (1654-1705), allows to describe binary phenomena where only events can occur with respective probabilities of p and $1-p$.

- Beta (a, b): the density function of this distribution (also called Beta type I) is given by:

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

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

We have $E(X) = \alpha/(\alpha+\beta)$ and $V(X) = \alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2]$

- Beta4 (α, β, c, d): the density function of this distribution is given by:

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

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

We have $E(X) = \frac{c+(c-d)\alpha}{(\alpha+\beta)}$ and $V(X) = \frac{(c-d)^2\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$

Pour the type I beta distribution, X takes values in the $[0,1]$ range. The beta4 distribution is obtained by a variable transformation such that the distribution is on a $[c,d]$ interval where c and d can take any value.

- Binomial (n,p) : the density function of this distribution is given by:

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

We have $E(X) = np$ and $V(X) = np(1-p)$

n is the number of trials, and p the probability of success. The binomial distribution is the distribution of the number of successes for n trials, given that the probability of success is p.

- Negative binomial type I (n,p): the density function of this distribution is given by:

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

We have $E(X) = n(1-p)/p$ and $V(X) = n(1-p)/p^2$

n is the number of successes, and p the probability of success. The negative binomial type I distribution is the distribution of the number x of unsuccessful trials necessary before obtaining n successes.

- Negative binomial type II (k,p): the density function of this distribution is given by:

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

We have $E(X) = kp$ and $V(X) = kp(p+1)$

The negative binomial type II distribution is used to represent discrete and highly heterogeneous phenomena. As k tends to infinity, the negative binomial type II distribution tends towards a Poisson distribution with $\lambda = kp$.

- $Khi^2(df)$: the density function of this distribution is given by:

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

We have $E(X) = df$ and $V(X) = 2df$

The Chi-square distribution corresponds to the distribution of the sum of df squared standard normal distributions. It is often used for testing hypotheses.

- Erlang (k, λ): the density function of this distribution is given by:

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

We have $E(X) = k/\lambda$ and $V(X) = k/\lambda^2$

k is the shape parameter and λ is the rate parameter.

This distribution, developed by the Danish scientist A. K. Erlang (1878-1929) when studying the telephone traffic, is more generally used in the study of queuing problems.

Note: When $k = 1$, this distribution is equivalent to the exponential distribution. The Gamma distribution with two parameters is a generalization of the Erlang distribution to the case where k is a real and not an integer (for the Gamma distribution the scale parameter $\beta = 1/\lambda$ is used).

- Exponential(λ): the density function of this distribution is given by:

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

We have $E(X) = 1/\lambda$ and $V(X) = 1/\lambda^2$

The exponential distribution is often used for studying lifetime in quality control.

- Fisher (df_1, df_2): the density function of this distribution is given by:

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

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

We have $E(X) = df_2/(df_2 - 2)$ if $df_2 > 2$, and $V(X) = \frac{2df_2^2(df_1+df_2-2)}{df_1(df_2-2)^2(df_2-4)}$ if $df_2 > 4$

Fisher's distribution, from the name of the biologist, geneticist and statistician Ronald Aylmer Fisher (1890-1962), corresponds to the ratio of two Chi-square distributions. It is often used for testing hypotheses.

- Fisher-Tippett (β, μ): the density function of this distribution is given by:

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

We have $E(X) = \mu + \beta\gamma$ and $V(X) = (\pi\beta)^2/6$ where γ is the Euler-Mascheroni constant.

The Fisher-Tippett distribution, also called the Log-Weibull or extreme value distribution, is used in the study of extreme phenomena. The Gumbel distribution is a special case of the Fisher-Tippett distribution where $\beta = 1$ and $\mu = 0$.

- Gamma (k, β, μ): the density of this distribution is given by:

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

We have $E(X) = \mu + k\beta$ and $V(X) = k\beta^2$

k is the shape parameter of the distribution and β the scale parameter.

- GEV (β, k, μ): the density function of this distribution is given by:

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

We have $E(X) = \mu + \frac{\beta}{k}\Gamma(1+k)$ and $V(X) = \left(\frac{\beta}{k}\right)^2 \left(\Gamma(1+2k) - \Gamma^2(1+k)\right)$

The GEV (Generalized Extreme Values) distribution is much used in hydrology for modeling flood phenomena. k lies typically between -0.6 and 0.6.

- Gumbel: the density function of this distribution is given by:

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

We have $E(X) = \gamma$ and $V(X) = \pi^2/6$ where γ is the Euler-Mascheroni constant (0.5772156649…).

The Gumbel distribution, named after Emil Julius Gumbel (1891-1966), is a special case of the Fisher-Tippett distribution with $\beta = 1$ and $\mu = 0$. It is used in the study of extreme phenomena such as precipitations, flooding and earthquakes.

- Logistic (μ, s): the density function of this distribution is given by:

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

We have $E(X) = \mu$ and $V(X) = (\pi s)^2/3$

- Lognormal (μ, σ): the density function of this distribution is given by:

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

We have $E(X) = exp(\mu + \sigma^2/2)$ and $V(X) = \left(exp(\sigma^2) - 1\right) exp(2\mu + \sigma^2)$

- Lognormal2 (m, s): the density function of this distribution is given by:

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

We have:

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

And:

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

This distribution is just a reparametrization of the Lognormal distribution.

- Normal (μ, σ) : the density function of this distribution is given by:

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

We have $E(X) = \mu$ and $V(X) = \sigma^2$

- Standard normal: the density function of this distribution is given by:

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

We have $E(X) = 0$ and $V(X) = 1$

This distribution is a special case of the normal distribution with $\mu = 0$ and $\sigma = 1$

- Pareto (a, b): the density function of this distribution is given by:

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

We have $E(X) = ab/(a-1)$ with $V(X) = \frac{ab^2}{((a-1)^2(a-2))}$

The Pareto distribution, named after the Italian economist Vilfredo Pareto (1848-1923), is also known as the Bradford distribution. This distribution was initially used to represent the distribution of wealth in society, with Pareto's principle that 80% of the wealth was owned by 20% of the population.

- PERT (a, m, b): the density function of this distribution is given by:

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

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

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

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

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

The PERT distribution is a special case of the beta4 distribution. It is defined by its definition interval [a, b] and m the most likely value (the mode). PERT is an acronym for *Program Evaluation and Review Technique*, a project management and planning methodology. The PERT methodology and distribution were developed during the project held by the US Navy and Lockheed between 1956 and 1960 to develop the Polaris missiles launched from submarines. The PERT distribution is useful to model the time that is likely to be spent by a team to finish a project. The simpler triangular distribution is similar to the PERT distribution in that it is also defined by an interval and a most likely value.

- Poisson (λ): the density function of this distribution is given by:

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

We have $E(X) = \lambda$ with $V(X) = \lambda$

Poisson's distribution, discovered by the mathematician and astronomer Siméon- Denis Poisson (1781-1840), pupil of Laplace, Lagrange and Legendre, is often used to study queuing phenomena.

- Student (df) : the density function of this distribution is given by:

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

We have $E(X) = 0$ if $df > 1$ with $V(X) = df/(df-2)$ if $df > 2$

The English chemist and statistician William Sealy Gosset (1876-1937), used the nickname Student to publish his work, in order to preserve his anonymity (the Guinness brewery forbade its employees to publish following the publication of confidential information by another researcher). The Student's t distribution is the distribution of the mean of df variables standard normal variables. When $df = 1$, Student's distribution is a Cauchy distribution with the particularity of having neither expectation nor variance.

- Trapezoidal (a, b, c, d): the density function of this distribution is given by:

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

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

This distribution is useful to represent a phenomenon for which we know that it can take values between two extreme values (a and d), but that it is more likely to take values between two values (b and c) within that interval.

- Triangular (a, m, b): the density function of this distribution is given by:

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

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

- TriangularQ (q_1, m, q_2, p_1, p_2): the density function of this distribution is a reparametrization of the Triangular distribution. A first step requires estimating the a and b parameters of the triangular distribution, from the q_1 and q_2 quantiles to which

percentages p_1 and p_2 correspond. Once this is done, the distribution functions can be computed using the triangular distribution functions.

- Uniform (a, b): the density function of this distribution is given by:

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

We have $E(X) = (a+b)/2$ with $V(X) = (b-a)^2/12$

The uniform (0,1) distribution is much used for simulations. As the cumulative distribution function of all the distributions is between 0 and 1, a sample taken in a Uniform (0,1) distribution is used to obtain random samples in all the distributions for which the inverse can be calculated.

- Uniform discrete (a, b): the density function of this distribution is given by:

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

We have $E(X) = (a+b)/2$ with $V(X) = [(b-a+1)^2 - 1]/12$

The uniform discrete distribution corresponds to the case where the uniform distribution is restricted to integers.

- Weibull (β): the density function of this distribution is given by:

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

We have $E(X) = \Gamma(\frac{1}{\beta} + 1)$ with $V(X) = \Gamma(\frac{2}{\beta} + 1) - \Gamma^2(\frac{1}{\beta} + 1)$

β is the shape parameter for the Weibull distribution.

- Weibull (β, γ): the density function of this distribution is given by:

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

We have $E(X) = \gamma\Gamma\left(\frac{1}{\beta} + 1\right)$ with $V(X) = \gamma^2\left[\Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right)\right]$

β is the shape parameter of the distribution and γ the scale parameter. When $\beta = 1$, the Weibull distribution is an exponential distribution with parameter $1/\gamma$.

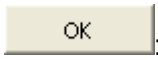
- Weibull (β, γ, μ): the density function of this distribution is given by:

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

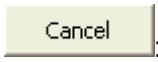
We have $E(X) = \mu + \gamma\Gamma\left(\frac{1}{\beta} + 1\right)$ with $V(X) = \gamma^2\left[\Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right)\right]$

The Weibull distribution, named after the Swede Ernst Hjalmar Waloddi Weibull (1887-1979), is much used in quality control and survival analysis. β is the shape parameter of the distribution and γ the scale parameter. When $\beta = 1$ and $\mu = 0$, the Weibull distribution is an exponential distribution with parameter $1/\gamma$.

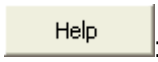
Dialog box



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

Theoretical distribution: Activate this option to sample data in a theoretical distribution. Then choose the distribution and enter any parameters required by the distribution.

Empirical Distribution: Activate this option to sample data in an empirical distribution. Then select the data required to build the empirical distribution.

Column labels: Activate this option if the first row of the selected data (data and weights) contains a label.

Weights: Activate this option if the observations are weighted. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column labels" option is activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Number of samples: Enter the number of samples to be generated.

Sample size: Enter the number of values to generate for each of the samples.

Display the report header: Deactivate this option if you want the table of sampled values to start from the first row of the Excel worksheet (situation after output to a worksheet or workbook) and not after the report header.

Example

An example showing how to generate a random normal sample is available on the XLSTAT Help Center:

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

References

Abramowitz M. & I.A. Stegun (1972). Handbook of Mathematical Functions. Dover Publications, New York, 925-964.

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.

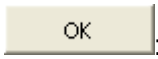
Variables transformation

Use this tool to quickly apply simple transformations to a set of variables.

In this section:

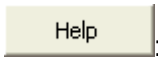
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.

: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

![Select_mouse.png][image-8]{: width="26" height="25"} ![Select_list.png][image-9]{: width="26" height="25"} ![Select_file.png][image-10]{: width="26" height="25"} : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files ![Select_file_choosefile.png][image-11]{: width="26" height="25"}.

General tab:

Data: Select the data in the Excel worksheet. If headers have been selected, check that the "Column labels" option has been activated.

Transformation: Choose the transformation to apply to the data.

- **Standardize (n-1)**: Choose this option to standardize the variables using the unbiased standard deviation.
- **Other**: Choose this option to use another transformation. Then click on the "Transformations" tab to choose the transformation to apply.
    - **Standardize (n)**: Choose this option to standardize the variables using the biased standard deviation.
    - **Center**: Choose this option to center the variables.
    - **/ Standard deviation (n-1)**: Choose this option to divide the variables by their unbiased standard deviation.
    - **/ Standard deviation (n)**: Choose this option to divide the variables by their biased standard deviation.
    - **Rescale from 0 to 1**: Choose this option to rescale the data from 0 to 1.
    - **Rescale from 0 to 100**: Choose this option to rescale the data from 0 to 100.
    - **Pareto scaling**: Choose this option to standardize the variables using the square root of the standard deviation.
    - **Binarize (0/1)**: Choose this option to convert all values that are not 0 to 1, and leave the 0s unchanged.
    - **Sign (-1/0/1)**: Choose this option to convert all values that are negative to -1, all positive values to 1, and leave the 0s unchanged.
    - **Arcsine**: Choose this option to transform the data to their arc-sine.
    - **Box-Cox transformation**: Activate this option to improve the normality of the sample; the Box-Cox transformation is defined by the following equation: $$ Y _ { t } = \left{ \begin{array} { l l } { \frac { X _ { t } ^ { \lambda } - 1} { \lambda } ,} & { (X _ { t } > 0,\lambda \neq 0) \text{ or } } (X _ { t } \geq 0,\lambda > 0) } \ { \operatorname{ln} (X _ { t }) ,} & { X _ { t } > 0,\lambda = 0} \end{array} \right. $$ XLSTAT accepts a fixed value of l, or it can find the value that maximizes the likelihood of the sample, assuming the transformed sample follows a normal distribution.
    - **Winsorize**: Choose this transformation to remove data that are not within an interval defined by two percentiles: let **p1** and **p2** be two values comprised between 0 and 1, such that $p1 < p2$. If a value x from the sample is lower than $q1$, the quantile that corresponds to $p1$ obtained from the sample, or greater than $q2$ the quantile that corresponds to $p2$, then the value is transformed to q1 in the first case, or to $q2$ in the second case.
    - **Johnson transformation**: Choose the Johnson transformation to transform your data to follow a normal distribution. This transformation is a generalization of the Box-Cox transformation which applies only to positive values. The selection of the

distribution and the estimation of the parameters is performed using the approach described by Chou *et al.* (1998):

- Johnson family S_B: $$ Y_{t} = \gamma + \eta ln(\frac{X_t - \epsilon}{\lambda + \epsilon - X_t}) $$ where $\eta, \lambda > 0, -\infty < \gamma < \infty, -\infty < \epsilon < \infty$ and $\epsilon < X_t < \epsilon + \lambda$

- Johnson family S_L: $$ Y_{t} = \gamma + \eta ln(X_t - \epsilon) $$ where $\eta > 0, -\infty < \gamma < \infty, -\infty < \epsilon < \infty$ and $\epsilon < X_t$

- Johnson family S_U: $$ Y_{t} = \gamma + \eta sinh^{-1}(\frac{X_t - \epsilon}{\lambda}) $$ where $\eta, \lambda > 0, -\infty < \gamma < \infty, -\infty < \epsilon < \infty$ and $-\infty < X_t < \infty$

You can select the normality test that is used to identify the best transformation together with the significance level.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the selected data (data and coding table) contains a label.

Display the report header: Deactivate this option if you want the results table to start from the first row of the Excel worksheet (situation after output to a worksheet or workbook) and not after the report header.

Observation labels: Check this option if you want to use the observation labels. If you do not check this option, labels will be created automatically (Obs1, Obs2, etc.). If a column header has been selected, check that the "Column labels" option has been activated.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations that contain missing data.

Ignore missing data: Activate this option to ignore missing data.

Estimate missing data: Activate this option to estimate missing data before starting computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

General formula: Activate this option to display the general formula that is used for the chosen transformation (except for the Johnson transformation).

Formula per variable: Activate this option to display the exact formula used by the chosen transformation (except for the "Binarize", "Sign" and "Arcsine" transformations).

Example

An example of a Johnson transformation is available on XLSTAT Help Center:

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

An example of a Box-Cox transformation is available on XLSTAT Help Center:

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

References

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.

Anonymizing data

Use this tool to anonymize your data. There are three possible methods: sequential, random and mapping. All three can be applied to both quantitative and qualitative data.

In this section:

Description

It's a good idea to transform sensitive private data before you share it. XLSTAT offers a tool to transform your quantitative and qualitative data according to 3 methods:

Sequential

The modalities of all selected variables are replaced by a sequentially selected integer starting at 1. The number associated with a modality appears as many times as the modality appears in the dataset, so the resulting file a numerical.

Random

This method varies depending on the variable type. For a quantitative variable, the values are randomly mixed. Thus they remain in the same scale as the initial data. For qualitative variables, the modalities are replaced by a string of randomly selected characters. This string is used as many times as the modality appears in the dataset.

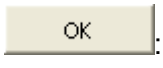
Mapping

Here the data is replaced by new values provided by a mapping table. This specifies the original value of the variables to be replaced in the left column and the new values in the right column.

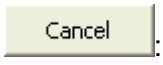
To complete the tool, you have the option of anonymizing variable labels.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. Here are descriptions of the various elements of the dialog box:



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display help options.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to samples. If the arrow points to the right, XLSTAT considers that rows correspond to samples and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files  .

General Tab:

Data: select the data you want to transform. If column headings have been selected, please make sure the “Variable labels” option is checked.

Anonymized variables: activate this option to anonymize variable labels.

Range: activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: activate this option to display the results in a new worksheet in the active workbook.

Workbook: activate this option to display the results in a new workbook.

Variable labels: activate this option if the first row of the selected data (data and observation labels) contains a label.

Observation labels: activate this option to display observation labels in the results if they are available. If this option is not activated, the observation labels are automatically generated by XLSTAT (Obs1, Obs2 …). If the "Variable labels" option is activated you must include a header in the selection.

Options Tab:

Anonymization methods: select how you want to anonymize your data from the 3 methods proposed: sequential, random and mapping.

Trim spaces: select this option so that XLSTAT deletes spaces before (check left) and/or after (check right) values in each selected cell.

Missing data tab:

Do not accept missing data: activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: activate this option to remove the observations with missing data.

Outputs Tab:

Original data: activate this option to display the table of selected data.

Anonymized data: activate this option to display the table of anonymized data.

Mapping table: activate this option to display the mapping table between selected and anonymized data.

Results

Original data: this table groups together all the selected data as displayed in the datasheet.

Anonymized data: this table groups together all the data that has been anonymized. They are displayed in the same order as the original data. If the option "Anonymized variables" has been chosen, then the variable labels are presented in their anonymous form.

Mapping table: this table lists the modalities of the variables that have been transformed. In the left column the initial values are listed, and in the right column the new ones.

Example

Check out this tutorial on anonymizing data with the XLSTAT Data-Anonymization module on the XLSTAT Help Center:

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

Missing data

Use this tool to handle missing values before running an analysis with XLSTAT.

In this section:

Description

Most XLSTAT features allows users to deal with missing data. However, only few approaches are available. This tool allows you to complete or clean your dataset using advanced techniques for missing data imputation.

There are three types of missing values (Allison, 2001): data missing completely at random (MCAR), data missing at random (MAR) and data not missing at random (NMAR).

Data missing completely at random (MCAR) occur if the event that leads to a missing data is independent of observable variables and unobservable parameters. It should happen entirely at random. When data are MCAR, the analyses performed on the data are unbiased.

Data missing at random (MAR) occur when the event that leads to a missing data is related to a particular variable, but it is not related to the value of the variable that has missing data. This is the most common case.

Data not missing at random (NMAR) occur when data is missing for a particular reason. An example of this is the filtered questions in a questionnaire (the question is only intended for some respondents)

The methods available in this tool correspond to the MCAR and MAR cases.

Different methods are available depending on the kind of problem and data:

- For quantitative data, XLSTAT allows to:
- Remove observations with missing value.
- Use a mean imputation method.
- Use a nearest neighbor approach.
- Replace missing values by a given numeric value.

- Use the NIPALS algorithm.
- Use an MCMC multiple imputation algorithm.
- Use the EM (Expectation Maximization) algorithm for data following a multivariate normal distribution.
- For qualitative data, XLSTAT allows to:
- Remove the observations with missing value.
- Use a mode imputation method.
- Use a nearest neighbor approach.
- Replace missing values by a given textual value.
- Use the NIPALS algorithm.

NIPALS algorithm

The NIPALS method is a method presented by H. Wold (1973) to allow principal component analysis with missing values. The NIPALS algorithm is applied on the dataset and the obtained PCA model is used to predict the missing values.

Multiple imputations

XLSTAT proposes a multiple imputation algorithm based on the Markov Chain Monte Carlo (MCMC) approach also called fully conditional specification (Van Buulen, 2007).

The algorithm works as follows:

1. Initial values of the missing values are obtained via a normal distribution sampling with mean and standard error equal to the mean and standard error obtained on available data.
2. For each variable of the dataset with missing values, an imputation method based on sampling and OLS regression is applied. The used model is a regression model with the studied variable as dependent variable and all the other variables as independent variables. Disturbance using data sampled from different distributions are also used. New imputed values are obtained using this model.

These two steps are repeated until the number of imputation is reached. The average value of each imputed missing value is taken.

EM algorithm

The EM algorithm used by XLSTAT must be applied to data following a **multivariate normal distribution**. Note that observations weights are not taken into account with this method.

The EM algorithm (Dempster, Laird, & Rubin, 1977) is a technique for estimating maximum likelihood (MLE) for incomplete data. For a detailed description of the applications of the EM algorithm, see the books of Little and Rubin (2002) and Schafer (1997).

The EM algorithm is an iterative procedure that finds the MLE of the multivariate normal distribution parameters by repeating the following steps:

1. Step E (Expectation): Given a set of parameters (mean vector and covariance matrix for a multivariate normal distribution), Step E calculates the expected likelihood of complete data based on observed data and parameter estimates.

2. Step M (Maximization): Based on the full data likelihood, Step M finds parameter estimates that maximize the full data likelihood obtained in Step E.

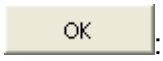
The two steps are repeated until convergence.

Remarks

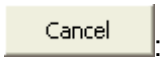
When you have quantitative variables and qualitative variables, and you choose to estimate the missing values, then the both tables are treated independently. On the other hand, if you choose to delete rows with missing values, then both tables will be merged and deleted rows on one table will also be deleted on the other table.

Dialog box

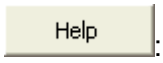
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Quantitative data: Activate this option to select a quantitative data table containing missing values. If a column header has been selected, check that the "Variable labels" option is activated.

Remove the observations: Activate this option to remove rows containing missing data.

Estimate missing data: Activate this option to impute missing data with the method of your choice (see Description section for more details on available methods).

- **Value**: If you chose to estimate all missing data with the same value, enter the numeric value to use.
- **Number of imputations**: If you selected the MCMC method, enter the number of imputations to be performed.
- **Iterations**: If you selected the EM algorithm, enter the number of iterations to perform.
- **Convergence**: If you selected the EM algorithm, enter the desired convergence level.

Qualitative data: Activate this option to select a qualitative data table containing missing values. If a column header has been selected, check that the "Variable labels" option is activated.

Remove the observations: Activate this option to remove rows containing missing data.

Estimate missing data: Activate this option to impute missing data with the method of your choice (see Description section for more details on available methods).

- **Value**: If you chose to estimate all missing data with the same value, enter the textual value to use.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the options in a new workbook.

Variable labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Weights: Activate this option if you want to carry out a weighted least squares regression. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected samples before and after imputation.

Missing values chart: Activate this option to display the missing values chart. This chart represents your dataset with the missing values colored in red. Variable labels also contain the percentage of missing data per column.

MCA results : Activate this option to display the proncipal coordinates table obtained by Multiple Correspondence Analysis (MCA) on missing data. For each variable, modality '0' represents the present data while modality '1' models the missing data. The map chart of this result is automatically displayed below.

Results

If you have checked all the proposed outputs then for each type of data (quantitative or qualitative) you will have in this order: - the missing values chart, - the descriptive statistics tables got before and after treatment, - the table of completed data whose imputed values are displayed in bold, - the ACM table and map graph of missing data for each variable.

Example

Examples showing how to run apply the NIPALS and EM imputation methods are available at:

http://www.xlstat.com/demo-missing.htm http://www.xlstat.com/demo-missing2.htm

References

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.

Raking a survey

Use this tool to compute raking weights using supplementary qualitative variables.

In this section:

Description

When working on surveys, one might need to rake the surveyed sample. That often means looking for raking weights. Sometime, a survey sample may cover a segment of the population with different proportions than in the population itself. The basic idea is to find weights applied to the observations that will give the same proportions in the survey sample as in the population. To do that, we need to use auxiliary variables that are measured both on the survey sample and on the population.

XLSTAT provides four methods for calculating adjustment weights, including the raking ratio method (Deming and Stefan, 1940).

Let's consider a sample of size n from a population with a known size N. A survey is conducted on this sample by including a certain number of auxiliary variables whose distribution over the global population is known. The adjustment methods will allow for iterative algorithms to find the appropriate weights so that the sample "looks like" the population.

The four methods available in XLSTAT are:

- The raking ratio method.
- The logit method: It is close to the raking ratio but lower and upper bounds for the weights are used. These bounds have to be specified by the user.
- The linear method.
- The truncated linear method: It is the linear method with bounds for the final weights.

The weights are calculated using an algorithm called the *generalized raking procedure* developed by Deville, Särndall and Sautory (1993). The difference from one method to another is the function to be optimized.

Raking weights computation

XLSTAT lets you use four methods to obtain raking weights. Classical raking ratio is one of them.

Let S be a sample with n observations and p qualitative variables called auxiliary. Let x_{ij} be the value of the i^{th} observation for j^{th} variable. Let X_j be the marginal control totals in the population for the j^{th} variable. Let d_i be the initial weights before raking and w_i the final weights after raking.

We are looking for weights w_i close to d_i satisfying the following equations:

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

We choose a distance function $G(u)$ with $u = w_k/d_k$, where this function has to be convex and positive. We have an optimization problem that can be solved using Lagrange multipliers method (λ).

The problem is the following:

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

With Lagrangian equal to:

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

We have:

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

With:

$$F(u) = g^{-1}(u) \quad \text{with } g(u) = \frac{dG(u)}{du}$$

A numerical method like Newton's method can be used to solve this problem.

The F() functions are:

- For the raking ratio method, we have:

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

- For the logit method, we have (with a lower bound L and an upper bound U):

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

- For the linear method, we have:

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

- For the truncated linear method, we have:

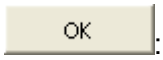
$$F(u) = 1 + u \in [L; U]$$

The algorithm is iterative and continues until the resulting weights are stable using the following criterion:

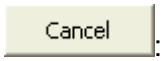
$$\max_{k=1}^{n} \left| \frac{w_k^{(i+1)}}{d_k} - \frac{w_k^{(i)}}{d_k} \right| < \epsilon$$

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data to be raked on: Select the auxiliary qualitative variables. If headers have been selected, check that the "Sample labels" option has been activated.

Marginal control totals: Select the marginal control totals for all the variables to be raked on. They should be represented in column with one row for each modality of the qualitative variable. Each column should sum to the same total. Columns have to be selected in the same orders as the variables to be raked on.

Format:

- **Values**: Activate this option if the marginal control totals represent real value in the global population.
- **Percentages**: Activate this option if the marginal control totals are percentages of the global population. If this option is selected, you have to indicate the **population size**.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Sample labels: Activate this option if the first row of the selected data (data, sub-samples, weights) contains a label.

Initial weights: Check this option if the observations are initially weighted. If you do not check this option, the weights will be considered as N/n. Weights must be greater than 0. If a column header has been selected, check that the "Sample labels" option is activated.

Observation labels: Activate this option if the observation labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observation labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Options tab:

Estimation method: Select the method you want to use. For logit and truncated linear, you have to give the bounds (lower bound < 1 and upper bound > 1).

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 50.
- **Convergence**: Enter the maximum value of the evolution in the convergence criterion from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.00001.

Missing data tab:

Remove observations: Activate this option to ignore the observations that contain missing data.

Estimate missing data: Activate this option to estimate the missing data by using the mode of the variables.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics of the sample before and after raking.

Show data in output: Activate this option to display the initial auxiliary variable in the final weights table.

Weights ratio: Activate this option to display the weights ratio for each observation in the final weights table.

List of combines: Activate this option to display the table with the list of all combines of the modalities and their associated frequency and weights ratio.

Details of iterations: Activate this option to display the table with the details for each iteration of the algorithm (Lagrange multipliers and stopping criterion).

Results

Summary statistics (before raking): This table displays for each modality of the auxiliary variables, the frequency and the percentages in the sample and in the population using marginal control totals.

Final weights: This table displays final raked weights. If the corresponding options are selected, initial data and weights ratios are also displayed.

Summary statistics (after raking): This table displays for each modality of the auxiliary variables, the frequency and the percentages in the sample with final weighting, and in the population using marginal control totals.

List of combines: This table displays all the combines of the auxiliary variable modalities with their frequency and their weights ratio.

Details of iterations: This table displays the details for each iteration with the Lagrange multipliers and the stopping criterion.

Example

An example showing how to rake a sample is available on the XLSTAT Help Center at:

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

References

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. (199 3). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, vol. 88, no. 418, 376-382.

Create a contingency table

Use this tool to create a contingency table from two or more qualitative variables. A chi-square test is optionally performed.

In this section:

Description

A contingency table is an efficient way to summarize the relation (or correspondence) between two categorical variables V1 and V2. It has the following structure:

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

where $n(i,j) = n_{ij}$ is the frequency of observations that show both characteristic i for variable V_1, and characteristic j for variable V_2.

To create a contingency table from two qualitative variables V_1 and V_2, the first transformation consists of recoding the two qualitative variables V_1 and V_2 as two disjunctive tables Z_1 and Z_2 or indicator (or dummy) variables. For each category of a variable there is a column in the respective disjunctive table. Each time the category c of variable V_1 occurs for an observation i, the value of $Z_1(i,c)$ is set to one (the same rule is applied to the V_2 variable). The other values of Z_1 and Z_2 are zero. The contingency table of the two variables is the table $Z_1' Z_2$ (where ' indicates matrix transpose).

The Chi-square distance has been suggested to measure the distance between two categories. The Pearson chi-square statistic, which is the sum of the Chi- square distances, is used to test the independence between rows and columns. Is has asymptotically a Chi-square distribution with $(m_1 - 1)(m_2 - 1)$ degrees of freedom.

Inertia is a measure inspired from physics that is often used in Correspondence Analysis, a method that is used to analyse in depth contingency tables. The inertia of a set of points is the weighted mean of the squared distances to the center of gravity. In the specific case of a

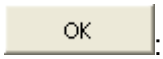
contingency table, the total inertia of the set of points (one point corresponds to one category) can be written as:

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

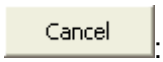
where n is the sum of the frequencies in the contingency table. We can see that the inertia is proportional to the Pearson chi-square statistic computed on the contingency table.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Row variable(s): Select the data that correspond to the variable(s) that will be used to construct the rows of the contingency table(s).

Column variable(s): Select the data that correspond to the variable(s) that will be used to construct the columns of the contingency table(s).

By group analysis: Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis on each group separately.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (row and column variables, weights) includes a header.

Weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Sort categories alphabetically: Activate this option so that the categories of all the variables are sorted alphabetically.

Variable-Category labels: Activate this option to create the labels of the contingency table using both the variable name of the name of the categories. If the option is not activated, the labels are only based on the categories.

Chi-square test: Activate this option to display the statistics and the interpretation of the Chi-square test of independence between rows and columns.

Significance level (%): Enter the significance level for the test.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations:

- **Pairwise deletion**: Activate this option to remove observations with missing data only when the variables involved in the calculations have missing data.
- **Across all Ys**: Choose this option to remove all observations with missing data.

Group missing values into a new category: Activate this option to group missing data into a new category of the corresponding variable.

Outputs tab:

List of combines: Activate this option to display the table that lists all the possible combines between the two variables that are used to create a contingency table, and the corresponding frequencies.

Contingency table: Activate this option to display the contingency table.

Inertia by cell: Activate this option to display the inertia for each cell of the contingency table.

Chi-square by cell: Activate this option to display the contribution to the chi-square of each cell of the contingency table.

Significance by cell: Activate this option to display a table indicating, for each cell, if the actual value is equal (=), lower (<) or higher (>) than the theoretical value, and to run a test (Fisher's exact test of on a 2×2 table having the same total frequency as the complete table, and the same marginal sums for the cell of interest), in order to determine if the difference with the theoretical value is significant or not. The associated p-values are also displayed.

Observed frequencies: Activate this option to display the table of the observed frequencies. This table is almost identical to the contingency table, except that the marginal sums are also displayed.

Theoretical frequencies: Activate this option to display the table of the theoretical frequencies computed using the marginal sums of the contingency table.

Proportions or percentages / Row: Activate this option to display the table of proportions or percentages computed by dividing the values of the contingency table by the marginal sums of each row.

Proportions or percentages / Column: Activate this option to display the table of proportions or percentages computed by dividing the values of the contingency table by the marginal sums of each column.

Proportions or percentages / Total: Activate this option to display the table of proportions or percentages computed by dividing the values of the contingency table by the sum of all the cells of the contingency table.

Summary across groups: Activate this option to display a summary of all contingency tables.

Charts tab:

3D view of the contingency table: Activate this option to display the 3D bar chart corresponding to the contingency table.

Contingency table: Activate this option to display the contingency table chart.

Proportions or percentages / Row: Activate this option to display the chart related to the *Proportions or percentages / Row* tab.

Proportions or percentages / Column: Activate this option to display the chart related to the *Proportions or percentages / Column* tab.

Summary across groups: Activate this option to display the charts associated with each of the groups in the summary table.

Chart options:

- **Chart type**
    - **Grouped**: Choose this option to display the graphs as bars grouped by modality.

- **Stacked bars**: Choose this option to display the chart as stacked bars. These charts are used to compare the frequencies of sub-samples to those of a full sample.

- **Bar charts**

  - **Frequencies**: Choose this option to display the frequencies corresponding to each bar.

  - **Percentages**: Choose this option to display the % of population corresponding to each bar.

Example

An example showing how to create a contingency table is available on the XLSTAT Help Center:

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

Full disjunctive tables

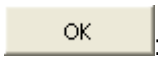
Use this tool to create a full disjunctive table from one or more qualitative variables.

In this section:

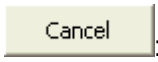
Description

A disjunctive table is a drill-down of a table defined by n observations and q qualitative variables into a table defined by n observations and p indicators where p is the sum of the numbers of categories of the q variables: each variable Q(j) is broken down into a sub-table with q(j) columns where column k contains 1's for observations corresponding to the k'th category and 0 for the other observations.

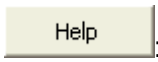
Dialog box



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

Data: Select the data in the Excel worksheet. If headers have been selected, check that the "Variable labels" option has been activated.

Variable labels: Check this option if the first line of the selected data contains a label.

Observation labels: Check this option if you want to use the available line labels. If you do not check this option, line labels will be created automatically (Obs1, Obs2, etc.). If a column header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

Display the report header: Deactivate this option if you want the full disjunctive table to start from the first row of the Excel worksheet (situation after output to a worksheet or workbook) and not after the report header.

Example

Input table:

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

Full disjunctive table:

	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

Multiple answer questions

Use this tool to transform tables that include multiple answer questions into a table where the multiple answer questions are transformed in such a way that they can be analyzed.

In this section:

Description

It is common that surveys included multiple answer questions. Here is very simple example: "Which are your favorite colors?". Some people will answer "Blue", other "Blue,Red", other "Green,Purple,Yellow". Any combination is possible from the list of possibilites that is givin in the survey. The output of the survey is a table giving the answers of each individual to each question. For multiple answer questions, it will typically be just as described above, with the different answers separated by a delimiter.

With this formatting, data analysis and statistical techniques requiring a structured layout cannot be applied. A transformation of the multple answer questions is necessary. The columns corresponding to the multiple answer questions are replaced by as many columns as there are answers in the input columns (Note: if an answer was never checked, a column will not be added for it), with within each column, Yes/No answers on whether the respondents have checked or not the item.

For the example above, we have:

|Colors| |--| |Blue| |Blue,Red| |Green,Purple,Yellow|	is transformed into	|Blue|Green|Purple|Red|Yellow| |--|--|--|--|--| |Yes|No|No|No|No| |Yes|No|No|Yes|No| |No|Yes|Yes|No|Yes|

Dialog box

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

Data: Select all the data that you want XLSTAT to analyse. If questions that are not multiple answer questions are included, XLSTAT will insert the columns corresponding to the different answers of the multiple answer questions, within the other questions. If column headers have been selected, check that the "Column labels" option has been activated.

Delimiter: Select the delimiter that is used to separate the multiple answers.

Confirm questions: Activate this option so that XLSTAT prompts you to confirm which questions are multiple answer questions in the data.

Results

XLSTAT displays the restructured table. If questions that are not multiple answer questions have been included, XLSTAT inserts the columns corresponding to the different answers of the multiple answer questions, within in between the other questions, respecting the order in the initial table.

Example

An example on how to use the tool is available on XLSTAT Help Center:

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

Discretization

Use this tool to discretize a numerical variable. Several discretization methods are available.

In this section:

Description

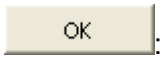
Discretizing a numerical variable means transforming it into an ordinal variable. This process is used a lot in marketing where it is often referred to as segmentation.

XLSTAT makes available several discretization methods that are more or less automatic. The number of classes (or intervals, or segments) to generate is either fixed by the user (for example with the method of equal ranges), or by the method itself (for example, with the 80-20 option where two classes are created).

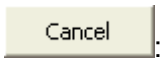
The Fisher's classification algorithm can be very slow when the size of dataset exceeds 1000. This method generates a number of classes that is lower or equal to the number of classes requested by the user, as the algorithm is able to automatically merge similar classes.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to samples. If the arrow points to the right, XLSTAT considers that rows correspond to samples and columns to observations.

General tab:

Observations/variables table: Select a table comprising N objects described by P descriptors. If column headers have been selected, check that the "Variable labels" option has been activated.

Method: Select the discretization method:

- **Constant range**: Choose this method to create classes that have the same range. Then enter the value of the range. You can optionally specify the "minimum" that corresponds to the lower bound of the first interval. This value must be lower or equal to the minimum value of the series. If the minimum is not specified, the lower bound will be set to the minimum value of the series.
- **Intervals**: Use this method to create a given number of intervals with the same range. The, enter the number of intervals. The range of the intervals is determined by the difference between the maximum and minimum values of the series. You can optionally specify the "minimum" that corresponds to the lower bound of the first interval. This value must be lower or equal to the minimum value of the series. If the minimum is not specified, the lower bound will be set to the minimum value of the series.
- **Equal frequencies**: Choose this method so that all the classes contain as much as possible the same number of observations. Then, enter the number of intervals (or classes) to generate.
- **Automatic (Fisher)**: Use this method to create the classes using the Fisher's algorithm. When the size of dataset exceeds 1000, the computations can be very slow. You need to enter the number of intervals (or classes) to generate. However; this method generates a number of classes that is lower or equal to the number of classes required by the user, as the algorithm is able to automatically merge similar classes.
- **Automatic (k-means)**: Choose this method to create classes (or intervals) using the k-means algorithm. Then, enter the number of intervals (or classes) to generate.
- **Intervals (user defined)**: Choose this option to select a column containing in increasing order the lower bound of the first interval, and the upper bound of all the intervals.
- **80-20**: Use this method to create two classes, the first containing the 80 first % of the series, the data being sorted in increasing order, the second containing the remaining 20%.
- **20-80**: Use this method to create two classes, the first containing the 20 first % of the series, the data being sorted in increasing order, the second containing the remaining

80%.

- **80-15-5 (ABC)**: Use this method to create two classes, the first containing the 80 first % of the series, the data being sorted in increasing order, the second containing the next 15%, and the third containing the remaining 5%. This method is sometimes referred to as "ABC classification".
- **5-15-80**: Use this method to create two classes, the first containing the 5 first % of the series, the data being sorted in increasing order, the second containing the next 15%, and the third containing the remaining 80%.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet in the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selected data contains a label.

Observation labels: Check this option if you want to use the available line labels. If you do not check this option, line labels will be created automatically (Obs1, Obs2, etc.). If a column header has been selected, check that the "Variable labels" option has been activated.

Display the report header: Deactivate this option if you do not want to display the report header.

Options tab:

Weights: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

- **Standardize the weights**: if you check this option, the weights are standardized such that their sum equals the number of observations.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations:

- **For the corresponding sample**: Activate this option to ignore an observation which has a missing value only for the variables that have a missing value.

- **For all samples**: Activate this option to ignore an observation which has a missing value for all selected variables.

Estimate missing data: Activate this option to estimate the missing data by using the mean of the variable.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Centroids: Activate this option to display the table of centroids of the classes.

Central objects: Activate this option to display the coordinates of the nearest object to the centroid for each class.

Results by class: Activate this option to display a table giving the statistics and the objects for each of the classes.

Results by object: Activate this option to display a table giving the class each object is assigned to in the initial object order.

Charts tab:

Histograms: Activate this option to display the histograms of the samples. For a theoretical distribution, the density function is displayed.

- **Bars**: Choose this option to display the histograms with a bar for each interval.
- **Continuous lines**: Choose this option to display the histograms with a continuous line.

Cumulative histograms: Activate this option to display the cumulated histograms of the samples.

- **Based on the histogram**: Choose this option to display cumulative histograms based on the same interval definition as the histograms.
- **Empirical cumulative distribution**: Choose this option to display cumulative histograms which actually correspond to the empirical cumulative distribution of the sample.

Ordinate of the histograms: Choose the quantity to be used for the histograms: density, frequency or relative frequency.

Results

Summary statistics: This table displays for the selected variables, the number of observations, the number of missing values, the number of non- missing values, the mean and the standard deviation.

A **histogram and the corresponding empirical cumulative distribution function** are displayed if the corresponding options are activated. The statistics of the intervals are then displayed.

Class centroids: This table shows the class centroids for the various descriptors.

Distance between the class centroids: This table shows the Euclidean distances between the class centroids for the various descriptors.

Central objects: This table shows the coordinates of the nearest object to the centroid for each class.

Distance between the central objects: This table shows the Euclidean distances between the class central objects for the various descriptors.

Results by class: The descriptive statistics for the classes (number of objects, sum of weights, within-class variance, minimum distance to the centroid, maximum distance to the centroid, mean distance to the centroid) are displayed in the first part of the table. The second part shows the objects.

Results by object: This table shows the assignment class for each object in the initial object order.

Example

An example showing how to discretize data is available on XLSTAT Help Center:

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

References

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.

Data management

Use this tool to manage tables of data. Eight functions are included in this tool: deduping, grouping, joining (inner and outer), filtering (keep and remove), stacking (and unstacking). These features are common in databases, but are not included in Excel.

In this section:

Description

Deduping

It is sometimes necessary to dedupe a table. Some observations might be mistakenly duplicated (or repeated) when they come from different sources, or because of input errors.

Grouping

Grouping is useful when you want to aggregate data. For example, imagine a table that contains all your sales records (one column with the customer id, and one with the sales value), that you want to transform to obtain one record per customer and the corresponding sum of sales. XLSTAT allows you to aggregate the data and to obtain the summary table within seconds. The sum is only one of the several available possibilities.

Joining

Joining is common task in database management. It allows to merge two tables "horizontally" on the basis of a common information named the "key". For example, imagine you measured some chemical indicators on 150 sites. Then you want to add geographical information on the sites where the data were collected. Your geographical table contains information on 1000 sites, including the 150 sites of interest. In order to avoid the tedious work of manually merging the two tables, a join will allow you to obtain within seconds the merged table that includes both the collected data and the geographical information.

One distinguishes two main types of joining:

- Inner joins: the merged table includes only keys that are common to both input tables.
- Outer joins: the merged table includes all keys that are available in the first, the second or both input tables.

Filtering (Keep/Remove)

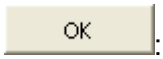
This tool allows you to select a table and create a new table that includes (Keep) or excludes (Remove) the rows, for which the value in a given column, matches a value contained in a user defined list.

Stack / Unstack

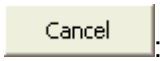
This tool allows you to transform a table organized in the form of a column by group in two columns, one associated with the value of the variable and the second to the associated group. It is also possible to stack only some of the selected columns, and keep the others intact. The reverse operation (unstack) is also possible. This allows for example to transform data in the form of columns into data suitable for a one-way analysis of variance.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to samples. If the arrow points to the right, XLSTAT considers that rows correspond to samples and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Data: This field is displayed if the selected method is "Dedupe", "Group","Filter" or "Stack". Select the data that correspond to the table that you want to dedupe, to aggregate, to filter or to stack.

Observation labels: This field is displayed only for the "Dedupe" method. Select the column (column mode) or row (row mode) where the observations labels are available. If you do not check this option, labels will be created automatically (Obs1, Obs2, etc.). If a column header has been selected, check that the "Variable labels" option has been activated.

Table 1: This field is displayed if the data management method is "Join". Select the data that correspond to the first input table to use in the join procedure.

Table 2: This field is displayed if the data management method is "Join". Select the data that correspond to the second input table to use in the join procedure.

Guess types: this option is displayed only for the "Group" method. Activate this option if you want that XLSTAT guesses the types of the variables of the selected table. If you uncheck this option, XLSTAT will prompt you to confirm or modify the type of the variables.

Method: select the data management method to use:

- Dedupe
- Group
- Join (Inner)
- Join (Outer)
- Filter (Keep)
- Filter (Remove)
- Stack
- Unstack

Operation: This option is only available if the method is "Group". Select the operation to apply to the data when aggregating them.

Stack certain columns: This option is only visible if the "Stack" method is activated. Check this option if you want to stack only part of your data and keep the rest. If the option is checked, a second interface will pops up when the analysis is launched allowing you to select the variables you wish to stack.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first row of the selected data (data and observation labels) contains a label.

Display the report header: Deactivate this option if you do not want to display the report header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in

the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Outputs tab:

This tab is only displayed if the selected method is "Dedupe" or "Group".

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

The following options are only displayed if the selected method is "Dedupe":

Deduped table: Activate this option to display the deduped table.

- **Frequencies**: Activate this option to display in the last column of the deduped table, the frequencies of each observation in the input table (1 corresponds to non-repeated observations; values equal or greater than 2 correspond to duplicated observations).

Duplicates: Activate this option to display the duplicates that have been removed from the original table in order to obtain the deduped table.

Missing data tab:

This tab is only displayed if the selected method is "Group".

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Ignore missing data: Activate this option to ignore missing data.

Text data cleaning

Use this tool to trim spaces to the left and/or right of strings, correct space repetitions or replace strings.

In this section:

Dialog box

General tab:

Data source: Choose the source of your text among these two options: * **Worksheet**: Select the data on the worksheet (one row per document) * **Document files**: Select one or multiple text files (WINDOWS version) or a folder containing them (MAC version).

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Column labels: Check this option if the first row of the selected data contains a label.

Display the report header: Deactivate this option if you do not want to display the report header.

Options tab:

Trim spaces: Activate this option to trim spaces from text data.

- **Left**: Activate this option to remove spaces at the beginning of the strings.
- **Right**: Activate this option to remove spaces at the end of the strings.
- **Including ** : Activate this option to remove the '\ ' characters (ASCII 160) that are typical of texts imported from the internet.

Spaces between words: Activate this option to define the maximum number of spaces between two words.

Replace characters: Activate this option to replace some characters (or words) with some other characters (or words). * **Replace**: Select a column with the characters (or words) to find and replace. * **By**: Select a column with the characters (or words) of replacements.

NB: This is a coding table entered in two steps, make sure the two columns have the same number of rows. The replacements are done in the order of entry and the case is taken into account. If headers have been selected, please check that the "Column labels" option has been activated.

Results

The results are displayed at the desired location. The table contains the strings processed by the methods chosen. The coding table is displayed if the "Replace characters" option is activated.

Coding

Use this tool to code or recode a table into a new table, using a coding table that contains the initial values and the corresponding new codes.

In this section:

Dialog box

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

Data: Select the data in the Excel worksheet. If headers have been selected, check that the "Column labels" option has been activated.

Coding table: Select a two-column table that contains in the first column the initial values, and in the second column the codes that will replace the values. If headers have been selected, check that the "Column labels" option has been activated.

Column labels: Activate this option if the first row of the data selected (data and coding table) contains a label.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Display the report header: Deactivate this option if you want the results table to start from the first row of the Excel worksheet (situation after output to a worksheet or workbook) and not after

the report header.

Example

An example showing how to recode data is available on the XLSTAT Help Center:

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

Presence/absence coding

Use this tool to convert a table of lists (or attributes) into a table of presences/absences showing the frequencies of the various elements for each of the lists.

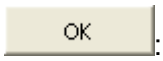
In this section:

Description

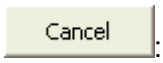
This tool is used, for example, to convert a table containing p columns corresponding to p lists of objects into a table with p rows and q columns where q is the number of different objects contained in the p lists, and where for each cell of the table, there is a 1 if the object is present and a 0 if it is absent.

For example, in ecology, if we have p species measurements with, for each measurement, the different species found in columns, we will obtain a two-way table showing the presence or absence of each of the species for each of the measurements.

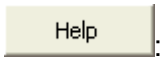
Dialog box



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

Data: Select the data in the Excel worksheet.

Column labels: Activate this option if the first row of the selected data contains a label.

Presence/absence coding by:

- **Rows**: Choose this option if each row corresponds to a list.

- **Columns**: Choose this option if each column corresponds to a list.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Display the report header: Deactivate this option if you want the results table to start from the first row of the Excel worksheet (situation after output to a worksheet or workbook) and not after the report header.

Example

Input table:

List 1	List 2
E1 E1 E2 E1 E3	E3 E1 E4

Presence/absence table:

	E1	E2	E3	E4
List 1	1	1	1	0
List 2	1	0	1	1

Coding by ranks

Use this tool to recode a table with n observations and p quantitative variables into a table containing ranks, the latter being determined variable by variable.

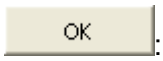
In this section:

Description

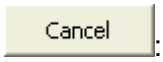
This tool is used to recode a table with n observations and p quantitative variables into a table containing ranks, the ranks being determined variable by variable. Coding in ranks lets you convert a table of continuous quantitative variables into discrete quantitative variables if only the order relationship is relevant and not the values themselves.

Two strategies are possible for taking tied values into account: either they are assigned to the mean rank or they are assigned to the lowest rank of the tied values.

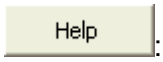
Dialog box



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

Data: Select the data in the Excel worksheet.

Variable labels: Check this option if the first line of the selected data contains a label.

Observation labels: Check this option if you want to use the available line labels. If you do not check this option, line labels will be created automatically (Obs1, Obs2, etc.). If a column header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Take ties into account: Activate this option to take account of the presence of tied values and adapt the rank of tied values as a consequence.

- **Mean ranks**: Choose this option to replace the rank of tied values by the mean of the ranks.
- **Minimum**: Choose this option to replace the rank of tied values by the minimum of their ranks.

Display the report header: Deactivate this option if you want the sampled table to start from the first row of the Excel worksheet (situation after output to a worksheet or workbook) and not after the report header.

Example

Initial table:

	V1	**V2**
Obs1	1.2	12
Obs2	1.6	11
Obs3	1.2	10
Obs4	1.4	10.5

Recoded table (using mean ranks for ties):

	R1	**R2**
Obs1	1.5	4
Obs2	4	3
Obs3	1.5	1
Obs4	3	2

Recoded table (using the lowest ranks for ties):

	R1	R2
Obs1	1	4
Obs2	4	3
Obs3	1	1
Obs4	3	2

Import data file

Use this tool to load data from a text file into a computer's memory. This method gives you the great advantage of using data which exceed the limits of an Excel worksheet.

In this section:

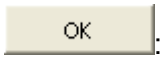
Description

Sometimes, it is necessary to work with data which have been stored into a text file. It can happen that this data cannot be imported into Excel due to volume constraints. To this end, XLSTAT allows to load data in a computer's memory (currently, this tool is only available for one functionnality, see Data management)

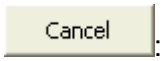
Dialog box

The dialog box is divided into two tabs including options ranging from data reading to file encoding. You will find below the description of the various elements of the dialog box.

: Click this button to show a preview of the first ten columns of the file.



: Click this button to save the parameters used to read the file.



: Click this button to close the dialog box without saving the parameters.


: Click this button to reload the default options.

General tab:

File path: this field defines the path where the data file is stored.

Format: select the file format:

- CSV Files: This includes all data files written with *.csv extension.
- Text Files: This includes all data files written with *.txt extension.
- All Files: This includes all data files written with all extension.

Delimiter: this field defines the charater used to separate each column in the file.

Decimal separator: this field defines the separator used for numerical values.

Encoding: select the encoding type:

- UTF-8: this encoding suggests that characters are written in 8 bits bytes. This fonctionnality checks if UTF-8 encoding uses a BOM (Byte Order Mark) and removes it if necessary.

- UTF-16: this encoding suggests that characters are written in 16 bits bytes. This fonctionnality reads BOM (Byte Order Mark) and defines if it's a big-endian (BE) or little-endian (LE) order. By default, if BOM doesn't exist, encoding is defined as a little-endian order (LE).

Text qualifier: this field defines the text qualifier. The data file can contain complex element such as elements composed by two words separated by a space. For example, SPACE chosen as delimiter and a column title is "Word list". If the text qualifier ' " ' is defined, "Word list" surrounded by quote will be see as a single element.

Variable labels: Check this option if the first row of the selected data (data and observation labels) contains a label.

Observation labels: Check this option if the first column of the selected data contains an observation label.

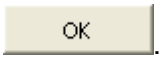
Options tab:

Start import at row: this field defines the row where XLSTAT starts reading the file. For example, you can choose not to include the firsts N lines of your data.

Comment qualifier: this field defines the character which is used commenting. All elements on the right of this character are ignored. Thus, a data file can containg comments without the these being considered in the data reading. For example, some MAC users have the habit to add a header with ' # ' character in their data in order to specify the generation date, author name, ...

Detect format: Check this option in order to automatically fill the parameters each time a **file path** is given.

Button **Preview** :

This button is used to show a preview window in which we can find every field of **General** tab plus the **Line(s) to show** field which allows you to define the number of rows displayed in the preview. It is possible to have a dynamic preview of the data while modifying the preview parameters. To validate these parameters, click on the button 

.

Describing data

Descriptive statistics and Univariate plots

Use this tool to calculate descriptive statistics and display univariate plots (Box plots, Scattergrams, etc) for a set of quantitative and/or qualitative variables.

In this section:

Description

Before using advanced analysis methods like, for example, discriminant analysis or multiple regression, you must first of all reveal the data in order to identify trends, locate anomalies or simply have available essential information such as the minimum, maximum or mean of a data sample.

XLSTAT offers you a large number of descriptive statistics and charts which give you a useful and relevant preview of your data.

Although you can select several variables (or samples) at the same time, XLSTAT calculates all the descriptive statistics for each of the samples independently.

Descriptive statistics for quantitative data:

Let's consider a sample made up of N items of quantitative data $\{y_1, y_2, \ldots y_N\}$ whose respective weights are $\{W_1, W_2, \ldots W_N\}$.

- **Number of observations**: The number N of values in the selected sample.
- **Number of missing values**: The number of missing values in the sample analyzed. In the subsequent statistical calculations, values identified as missing are ignored. We define **n** to be the number of non-missing values, and $\{x_1, x_2, \ldots x_n\}$ to be the sub-sample of non-missing values whose respective weights are $\{w_1, w_2, \ldots w_n\}$.
- **Sum of weights** *: The sum of the weights, **W**. When all weights are 1, or when weights are "standardized", W=n.

- **Breakdown per subsample (%)** *: The breakdown of each of the subsamples.
- **Minimum**: The minimum of the series analyzed.
- **Maximum**: The maximum of the series analyzed.
- **Frequency of minimum** *: The frequency of the minimum of the series.
- **Frequency of maximum** *: The frequency of the maximum of the series.
- **Range**: The range is the difference between the minimum and maximum of the series.
- **1st quartile** *: The first quartile **Q1** is defined as the value for which 25% of the values are less.
- **Median** *: The median **Q2** is the value for which 50% of the values are less.
- **3rd quartile** *: The third quartile **Q3** is defined as the value for which 75% of the values are less.
- **Sum** *: The weighted sum of the values is defined by:

$$S = \sum_{i=1}^{n} w_i x_i$$

- **Mean** *: The mean of the sample is defined by:

$$\hat{\mu} = S/W$$

- **Variance(n)** *: The variance of the sample defined by:

$$s_n^2 = \frac{\sum_{i=1}^{n} w_i (x_i - \hat{\mu})^2}{W}$$

Note 1: When all the weights are 1, the variance is the sum of the square deviation to the mean divided by n, hence its name.

Note 2: The variance (n) is a biased estimate of the variance which assumes that the sample is a good representation of the total population. The variance (n-1) is, on the other hand, calculated taking into account an approximation associated with the sampling.

- **Variance(n-1)** *: The estimated variance of the sample defined by:

$$s_{n-1}^2 = \frac{\sum_{i=1}^{n} w_i (x_i - \hat{\mu})^2}{W - W/n}$$

Note 1: When all the weights are 1, the variance is the sum of the square deviation to the mean divided by n-1, hence its name.

Note 2: The variance(n) is a biased estimate of the variance which assumes that the sample is a good representation of the total population. The variance(n-1) is, on the other hand, calculated taking into account an approximation associated with the sampling.

- **Standard deviation(n)** *: The standard deviation of the sample defined by $s_n = \sqrt{s_n^2}$.
- **Standard deviation(n-1)** *: The standard deviation of the sample defined by $s_{n-1} = \sqrt{s_{n-1}^2}$.
- **Variation coefficient** *: this coefficient is only calculated if the mean of the sample is non-zero. It is defined by:

$$CV = \frac{s_n}{\hat{\mu}}$$

This coefficient measures the dispersion of a sample relative to its mean. It is used to compare the dispersion of samples whose scales or means differ greatly.

- **Skewness (Pearson)** *: The sample Pearson skewness coefficient is defined by:

$$g_1 = \frac{\hat{\mu_3}}{s_{n-1}^3} \text{ with } \hat{\mu_3} = \frac{\sum_{i=1}^{n} w_i(x_i - \hat{\mu})^3}{W}$$

This coefficient gives an indication of the shape of the distribution of the sample. If the value is negative (or positive respectively), the distribution is concentrated on the left (or right respectively) of the mean.

- **Skewness (Fisher)** *: The Fisher skewness coefficient is defined by:

$$G_1 = g_1 \frac{\sqrt{W(W - W/n)}}{W - 2W/n}$$

Unlike the previous, this coefficient is not biased on the assumption that the data is normally distributed. This coefficient gives an indication of the shape of the distribution of the sample. If the value is negative (or positive respectively), the distribution is concentrated on the left (or right respectively) of the mean.

- **Skewness (Bowley)** *: The Bowley skewness coefficient is defined by:

$$A(B) = \frac{Q_1 - 2Q_2 + Q_3}{Q_3 - Q_1}$$

- **Kurtosis (Pearson or excess)**: The Fisher kurtosis coefficient is defined by:

$$g_2 = \frac{\hat{\mu_4}}{s_{n-1}^4} - 3 \text{ with } \hat{\mu_4} = \frac{\sum_{i=1}^{n} w_i(x_i - \hat{\mu})^4}{W}$$

This coefficient, sometimes called *excess kurtosis*, gives an indication of the shape of the distribution of the sample. If the value is negative (or positive respectively), the peak of the distribution of the sample is more flattened out (or respectively less) than that of a normal distribution.

- **Kurtosis (Population)** *: The unbiased kurtosis coefficient of a population is defined by:

$$G_2 = \frac{(W - W/n)}{(W - 2W/n)(W - 3W/n)}((W + W/n)g2 + 6)$$

Unlike the previous, this coefficient is not biased on the assumption that the data is normally distributed. This coefficient, sometimes called *excess kurtosis*, gives an indication of the shape of the distribution of the sample. If the value is negative (or positive respectively), the peak of the distribution of the sample is more flattened out (or respectively less) than that of a normal distribution.

- **Standard error of the mean** *: this statistic is defined by:

$$s_\mu = \frac{s_n}{\sqrt{n-1}}$$

- **Lower bound on mean (x% or significance level α=1-x/100)** *: this statistic corresponds to the lower bound of the confidence interval at x% of the mean. This statistic is defined by:

$$L_\mu = \hat{\mu} - s_\mu |t_{(a/2)}|$$

- **Upper bound on mean (x% or significance level α=1-x/100)** *: this statistic corresponds to the upper bound of the confidence interval at x% of the mean. This statistic is defined by:

$$U_\mu = \hat{\mu} + s_\mu |t_{(\alpha/2)}|$$

- **Standard error of the variance** *: this statistic is defined by:

$$s_{s^2} = s^2_{n-1}\sqrt{\frac{2}{W-1}}$$

- **Lower bound on mean (x% or significance level α=1-x/100)** *: this statistic corresponds to the lower bound of the confidence interval at x% of the variance. This statistic is defined by:

$$L_{s^2} = s_\sigma / \chi_{(1-\alpha/2)}$$

- **Upper bound on mean (x% or significance level α=1-x/100)** *: this statistic corresponds to the upper bound of the confidence interval at x% of the variance. This statistic is defined by:

$$U_{s^2} = s_\sigma / \chi_{(\alpha/2)}$$

- **Standard error (Skewness (Fisher))** *: The standard error of the Fisher's skewness coefficient is defined by:

$$se(G_1) = \sqrt{\frac{6W(W-1)}{(W-2)(W+1)(W+3)}}$$

- **Standard error (Kurtosis (Fisher))** *: The standard error of the Fisher's kurtosis coefficient is defined by:

$$se(G_2) = 2se(G1)\sqrt{\frac{(W^2-1)}{(W-3)(W+5)}}$$

- **Mean absolute deviation** *: as for standard deviation or variance, this coefficient measures the dispersion (or variability) of the sample. It is defined by:

$$e(\mu) = \frac{\sum_{i=1}^{n} w_i|x_i - \mu|}{W}$$

- **Median absolute deviation**: this statistic is the median of absolute deviations to the median.

- **Geometric mean**: this statistic is only calculated if all the values are strictly positive. It is defined by:

$$\mu_G = exp(\frac{1}{W}\sum_{i=1}^{n} w_i ln(x_i))$$

If all the weights are equal to 1, we have:

$$\mu_G = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

- **Geometric standard deviation**: this statistic is defined by:

$$\sigma_G = exp(\frac{1}{W}\sum_{i=1}^{n} w_i [ln(x_i) - ln(\mu_G)]^2)$$

- **Harmonic mean**: this statistic is defined by:

$$\mu_H = \frac{W}{\sum_{i=1}^{n} \frac{w_i}{x_i}}$$

(*) Statistics followed by an asterisk take the weight of observations into account.

Descriptive statistics for qualitative data:

For a sample made up of N qualitative values, we define:

- **Number of observations**: The number N of values in the selected sample.

- **Number of missing values**: The number of missing values in the sample analyzed. In the subsequent statistical calculations, values identified as missing are ignored. We define n to be the number of non-missing values, and {w1, w2, … wn} to be the sub-sample of weights for the non-missing values.

- **Sum of weights***: The sum of the weights, **W**. When all the weights are 1, W=n.

- **Mode** *: The mode of the sample analyzed. In other words, the most frequent category.
- **Frequency of mode** *: The frequency of the category to which the mode corresponds.
- **Category**: The names of the various categories present in the sample.
- **Frequency by category** *: The frequency of each of the categories.
- **Relative frequency by category** *: The relative frequency of each of the categories.
- **Lower bound on frequencies (x% or significance level α=1-x/100)** *: This statistic corresponds to the lower bound of the confidence interval at x% of the frequency per category.
- **Upper bound on frequencies (x% or significance level α=1-x/100)** *: This statistic corresponds to the upper bound of the confidence interval at x% of the frequency per category.
- **Proportion per category***: The proportion of each of the categories.
- **Lower bound on proportions (x% or significance level α=1-x/100)** *: This statistic corresponds to the lower bound of the confidence interval at x% of the proportion per category.
- **Upper bound on proportions (x% or significance level α=1-x/100)** *: This statistic corresponds to the upper bound of the confidence interval at x% of the proportion per category.

(*) Statistics followed by an asterisk take the weight of observations into account.

Several types of chart are available for quantitative and qualitative data:

Charts for quantitative data:

- **Box plots**: These univariate representations of quantitative data samples are sometimes called "box and whisker diagrams". It is a simple and quite complete representation since in the version provided by XLSTAT the minimum, 1-st quartile, median, mean and 3-rd quartile are displayed together with both limits (the ends of the "whiskers") beyond which values are considered anomalous. The mean is displayed with a red +, and a black line corresponds to the median. Limits are calculated as follows:

Lower limit: $L_{inf} = X(i)$ such that $\{X(i) - [Q1 - 1.5(Q3 - Q1)]\}$ is minimum and $X(i) = Q1 - 1.5(Q3 - Q1)$.

Upper limit: $L_{sup} = X(i)$ such that $\{X(i) - [Q3 + 1.5(Q3 - Q1)]\}$ is minimum and $X(i) = Q3 + 1.5(Q3 - Q1)$

Values that are outside the $]Q1-3(Q3-Q1); Q3+3(Q3-Q1)[$ interval are displayed with the * symbol; values that are in the $[Q1-3(Q3-Q1); Q1-1.5(Q3-Q1)]$ or the $[Q3+1.5(Q3-Q1); Q3+3(Q3-Q1)]$ intervals are displayed with the "o" symbol.

XLSTAT allows producing "notched" box plots. The limits of the notch allow to visualize a 95% confidence interval around the median. The limits are given by:

- **Lower limit**: $N_{inf} = Median - [1.58(Q3-Q1)]/\sqrt{(n)}$
- **Upper limit**: $N_{sup} = Median + [1.58(Q3-Q1)]/\sqrt{(n)}$

These formulae given by McGill *et al.* (1978) derive from the asumption that the medians are normally distributed and coming from equal size samples. If the sample sizes are indeed similar notched box plots allow to tell whether the samples have different medians or not and to compare their variability using the size of the notch.

XLSTAT allows to make the box plots width vary with the sample size. The width is proportional to the square root of the sample size.

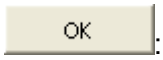
- **Scattergrams**: These univariate representations give an idea of the distribution and possible plurality of the modes of a sample. All points are represented together with the mean and the median.
- **Strip plots**: These diagrams represent the data from the sample as strips. For a given interval, the thicker or more tightly packed the strips, the more data there is.
- **Stem-and-leaf plots**: These univariate representations help to visualize the data distribution while keeping a precise view over the values, in opposition to histograms. Each data point is split into two parts: a stem part, left of the diagram, and a leaf part, right of the diagram. Each data point can be recovered by multiplying the number [stem.leaf] by the unit displayed at the top of the representation.
- **P-P Charts (normal distribution)**: P-P charts (for Probability-Probability) are used to compare the empirical distribution function of a sample with that of a normal variable for the same mean and deviation. If the sample follows a normal distribution, the data will lie along the first bisector of the plan.
- **Q-Q Charts (normal distribution)**: Q-Q charts (for Quantile-Quantile) are used to compare the quantities of the sample with that of a normal variable for the same mean and deviation. If the sample follows a normal distribution, the data will lie along the first bisector of the plan.
- **Means charts**: The means charts represent, in the form of a bar chart, the means of each of the variables. It is also possible to display the **error bars** on these graphs in three different forms:
    - the standard deviation defined by $s_n = \sqrt{s_n^2}$.
    - the standard error defined by $err = \frac{s_n}{\sqrt{n}}$.
    - the confidence interval defined by $L_\mu = \hat{\mu} \pm s_\mu |t_{(a/2)}|$.

Charts for qualitative data:

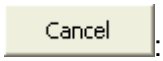
- **Bar charts**: Check this option to represent the frequencies or relative frequencies of the various categories of qualitative variables as bars.
- **Pie charts**: Check this option to represent the frequencies or relative frequencies of the various categories of qualitative variables as pie charts.
- **Double pie charts**: These charts are used to compare the frequencies or relative frequencies of sub-samples with those of the complete sample.
- **Doughnuts**: This option is only checked if a column of sub-samples has been selected. These charts are used to compare the frequencies or relative frequencies of sub-samples with those of the complete sample.
- **Stacked bars**: This option is only checked if a column of sub-samples has been selected. These charts are used to compare the frequencies or relative frequencies of sub-samples with those of the complete sample.
- **Multiple bars**: This option is only checked if a column of sub-samples has been selected. These charts are used to compare the frequencies or relative frequencies of sub-samples with those of the complete sample.

Dialog box

The dialog box is made up of several tabs corresponding to the various options for controlling the calculations and displaying the results. A description of the various components of the dialog box are given below.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Quantitative data: Check this option to select the samples of quantitative data you want to calculate descriptive statistics for.

Qualitative data: Check this option to select the samples of qualitative data you want to calculate descriptive statistics for.

Subsamples: Check this option to select a column showing the names or indexes of the sub-samples for each of the observations.

- **Variable-Category labels**: Activate this option to use variable-category labels when displaying outputs. Variable-Category labels include the variable name as a prefix and the category name as a suffix.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Sample labels: Check this option if the first line of the selections (quantitative data, qualitative date, sub-samples, and weights) contains a label.

Weights: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Sample labels" option is activated.

- **Standardize the weights**: if you check this option, the weights are standardized such that their sum equals the number of observations.

Options tab:

Descriptive statistics: Check this option to calculate and display descriptive statistics.

Charts: Check this option to display the charts.

Normalize: Check this option to standardize the data before carrying out the analysis.

Rescale from 0 to 100: Check this option to arrange the data on a scale of 0 to 100.

Compare to total sample: this option is only checked if a column of sub- samples has been selected. Check this option so that the descriptive statistics and charts are also displayed for the total sample.

Confidence interval: Enter the size of the confidence interval (in %).

Outputs tab:

Quantitative Data: Activate the options for the descriptive statistics you want to calculate. The various statistics are described in the description section.

All: Click this button to select all.

None: Click this button to deselect all.

- **Display vertically**: Check this option so that the table of descriptive statistics is displayed vertically (one line per descriptive statistic).

Qualitative Data: Activate the options for the descriptive statistics you want to calculate. The various statistics are described in the description section.

All: Click this button to select all.

None: Click this button to deselect all.

- **Display vertically**: Check this option so that the table of descriptive statistics is displayed vertically (one line per descriptive statistic).

Charts (1) tab:

This tab deals with the quantitative data.

Chart types sub-tab:

Box plots: Check this option to display box plots (or box-and-whisker plots). See the description section for more details.

Scattergrams: Check this option to display scattergrams. The mean (red +) and the median (red line) are always displayed.

Strip plots: Check this option to display strip plots. On these charts, a strip corresponds to an observation.

Stem-and-leaf plots: Check this option to display stem-and-leaf plots.

- **Unit: 10^**: Activate this option if you wish to configure manually the unit over which data points will be split into stems and leaves.

Normal P-P plots: Check this option to display P-P plots.

Normal Q-Q Charts: Check this option to display Q-Q plots.

Means charts: check this option to display the means charts.

Options sub-tab:

These options concern box plots, scattergrams and strip plots

Horizontal: Check this option to display box plots, scattergrams and strip plots horizontally.

Vertical: Check this option to display box plots, scattergrams and strip plots vertically.

Group plots: Check this option to group together the various box plots, scattergrams and strip plots on the same chart to compare them.

It is possible to specify the "**Dimension**" corresponding to the maximum number of boxplots to group. By default, this number is **automatically** chosen based on the number of variables and categories. You can also specify it manually. This number can be at most 20 for Excel 2003 and 40 for Excel 2007 and beyond.

- **Categories**: Check this option if you want to group all categories per variable. This option is available when the Subsamples field (general tab) is used.
- **Variables**: Check this option if you want to group all variables per category. This option is available when the Subsamples field (general tab) is used. Check the **Grey line** option to separate variables with grey lines on the plots.
- **Sort by mean**: Check this option to sort the variables or categories by decreasing mean.

Notched: Check this option if you want to display notched box plots.

Adapt width: Check this option if you want that the width of the box plots depends on the sample size.

Minimum/Maximum: Check this option to systematically display the points corresponding to the minimum and maximum (box plots).

Outliers: Check this option to display the points corresponding to outliers (box plots) with a hollowed-out circle.

Labels position: Select the position where the labels have to be placed on the box plots, scattergrams and strip plots.

Legend: Activate this option to display the legend describing the statistics used on the box plot.

Color inside: Activate this option to color the inside of the box plots.

Color by group: Activate this option so that the color of the box plots varies for each subsample, if subsamples have been selected.

Charts (2) tab:

This tab deals with the qualitative data.

Bar charts: Check this option to represent the frequencies or relative frequencies of the various categories of qualitative variables as bars.

Pie charts: Check this option to represent the frequencies or relative frequencies of the various categories of qualitative variables as pie charts.

- **Doubles**: this option is only checked if a column of sub-samples has been selected. These charts are used to compare the frequencies or relative frequencies of sub-samples with those of the complete sample.

Doughnuts: this option is only checked if a column of sub-samples has been selected. These charts are used to compare the frequencies or relative frequencies of sub-samples with those of the complete sample.

Stacked bars: this option is only checked if a column of sub-samples has been selected. These charts are used to compare the frequencies or relative frequencies of the different groups of the sub-sample.

Multiple bars: this option is only checked if a column of sub-samples has been selected. These charts are used to compare the frequencies or relative frequencies of the different groups of the sub-sample.

Values used: choose the type of data to be displayed:

- **Frequencies**: choose this option to make the scale of the plots correspond to the frequencies of the categories.
- **Relative frequencies**: choose this option to make the scale of the plots correspond to the relative frequencies of the categories.

Example

An example showing how to create Box plots is available on XLSTAT Help Center:

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

References

Filliben J.J. (1975). The Probability Plot Correlation Coefficient Test for Normality. Technometrics, 17(1), pp 111-117.

Lawrence T. DeCarlo (1997). On the Meaning and Use of Kurtosis. Psychological Methods, 2(3), pp. 292-307.

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.

Variable characterization

Use this tool to characterize elements (quantitative variables, qualitative variables or categories of qualitative variables) exploring the links they share with characterizing elements (quantitative variables, qualitative variables or categories of qualitative variables). For this purpose, different statistical tests (parametric or non-parametric) are used.

In this section:

Description

Here are the various characterizing possibilities of the procedure:

1. Characterization of a quantitative variable:

1-1. With other quantitative variables:

Characterization of a quantitative variable by other quantitative variables is done with the correlation coefficient. For each characterizing quantitative variable, it is tested whether the correlation coefficient is significantly different from 0. Pearson correlation coefficient is used in the case of parametric test and Spearman correlation coefficient is used in the non-parametric case. The more significantly the correlation coefficient differs from 0, the more the two quantitative variables are linked.

1-2. With qualitative variables:

Characterization of a quantitative variable by qualitative variables is performed using parametric or non-parametric statistical tests. If the p-value of the test is lower than a selected threshold the assumption of independence between the two variables is rejected. In the parametric case Fisher test is used (as in ANOVA). In the non-parametric case if the qualitative variable has $k = 2$ categories, Mann-Whitney test is used, if the qualitative variable has more than 2 categories, Kruskal-Wallis test is used.

1-3. With categories of a qualitative variable:

Characterization of a quantitative variable with categories of a qualitative variable is done using an indicator called test value (Lebart, 2000). The test value $t_k(X)$ of a quantitative variable X associated with the category k of a qualitative variable is defined as follows:

$$t_k(X) = \frac{\overline{X_k} - \overline{X}}{s_k(X)}$$

with :

$$s_k^2(X) = \frac{n - n_k}{n - 1} \frac{s^2(X)}{n_k}$$

where $s^2(X)$ is the empirical variance of the variable X. A p-value associated with this test value is then calculated, the closer the p-value is to 0, the more the average of the variable X on the category k is different from the general average.

1. Characterization of a qualitative variable (with k categories):

2-1. With quantitative variables:

Characterization of a qualitative variable by quantitative variables is performed using parametric or non-parametric statistical tests. If the p-value of the test is lower than a selected threshold the assumption of independence between the two variables is rejected. In the parametric case, the Fisher test is used (as in ANOVA). In the non-parametric case if the qualitative variable has $k = 2$ categories, the Mann-Whitney test is used, if the qualitative variable has more than 2 categories, Kruskal-Wallis test is used.

2-2. With other qualitative variables:

Characterization of a qualitative variable by other qualitative variables is carried out using an independence test. For each characterizing qualitative variable we test the independence with the qualitative variable to characterize with the Chi² independence test (parametric) or the exact Fisher test (nonparametric).

1. Characterization of a category of a qualitative variable:

3-1. With quantitative variables:

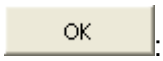
Characterization of a category of a qualitative variable by quantitative variables is done using the test value as explained in 1-3.

3-2. With categories of a qualitative variable:

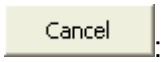
Characterization of a category with other categories is done using the test value for qualitative variables (Lebart, 2000) and its associated p-value. A category is considered to be characterizing a class if its abundance in the class is considered significantly superior to what can be expected given its presence in the whole population. Let n_{jk} be the number of individuals with the category j among the n_k individuals in the class k, n_j the number of individuals with the category j and n the total number of individuals, the abundance of the category j is defined by comparing its percentage in the kth class $\frac{n_{kj}}{n_k}$ to its percentage in the population $\frac{n_j}{n}$.

Dialog box

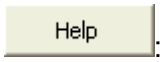
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Element(s) to characterize:

Quantitative variable(s): Activate this option if you want to characterize one or several quantitative variables. Then, select the variable(s) you want to characterize. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

Qualitative variable(s): Activate this option if you want to characterize one or several qualitative variables. Then, select the variable(s) to characterize. If several variables have been selected, XLSTAT carries out calculations for each variable separately. If a column header has been selected, check that the "Variable labels" option has been activated.

Categories: Activate this option if you want to characterize the categories of the qualitative variable(s) previously selected.

X / Characterizing elements:

Quantitative Variable(s): Activate this option if you want characterizing quantitative variable(s). Then, select the quantitative variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative Variable(s): Activate this option if you want characterizing qualitative variable(s). Then, select the qualitative variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Categories: Activate this option if you want to use the categories of the qualitative variable(s) previously selected as characterizing elements.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selections (quantitative data, qualitative date, and weights) contains a label.

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. In the case of non-parametric tests, weights must be positive integers. If a column header has been selected, check that the "Variable labels" option has been activated.

Options tab:

Filter characterizing elements: Activate this option if you want to filter the characterizing elements to display. Several filtering options are available, depending on the chosen option, you must choose a threshold for the p-values (or test values) to display or a number p of characterizing elements to display.

Sort characterizing elements: Activate this option if you want to sort the display of the characterizing elements according to the p-values.

Significance level: Enter the significance level you want in the associated cell.

Parametric tests: Active this option if you want to perform a parametric test.

Non-parametric tests: Active this option if you want to perform a non-parametric test.

Missing data tab:

Remove observations: Activate this option to remove an observation that has a missing value.

Estimate missing data: Activate this option to estimate the missing data by using the mean of the sample or the nearest neighbor.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

p-value chart: Activate this option to display the chart of p-values.

Test values chart: Activate this option to display the chart of test values.

Results

Summary statistics: This table displays descriptive statistics for all the variables selected.

The result table differs depending on the elements to be characterized as well as the characterizing elements. In all cases, p-values will always be displayed.

If you have selected the **p-value chart** option, a bar chart with p-values is also displayed below each table.

Example

An example showing how to compute variable characterization is available on XLSTAT Help Center:

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

References

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** , n° 1-2, 20-27.

Quantiles estimation

Use this tool to calculate quantiles and display univariate plots (Box plots, Scattergrams, etc) for a set of quantitative variables.

In this section:

Description

Quantiles (or percentiles) can be very useful in statistics. A percentile is a quantile based on a 0 to 100 scale.

XLSTAT offers you five methods to calculate quantiles. Furthermore, two types of confidence intervals are available.

While you can select several samples at the same time, XLSTAT calculates all the descriptive statistics for each sample independently.

Definition of a quantile

Let $0 < p < 1$. The p-quantile of a variable X is given by:

$$P(X \leq x) \geq p \text{ and } P(X \geq x) \geq 1 - p$$

Quantiles are useful because they are less sensitive to outliers and skewed distributions.

Methods for quantile computation

Five different methods are available in XLSTAT. Let's consider a sample made up of N items of quantitative data $\{x_1, x_2, \ldots x_N\}$ whose respective weights are $\{w_1, w_2, \ldots w_N\}$. Let $x(1), \ldots, x(N)$ be the ordered data.

Let y be the p-quantile, j be the integer part of $N \times p$ and g be the fractional part. We have: $g = N \times p - j$

We have:

Weighted average at $x(N*p)$:

$$y = (1-g)x_{(j)} + gx_{(j+1)}$$

where x(0) is replaced by x(1).

Observation numbered closest to $x(N*p)$:

$$\begin{aligned} &y = x_{(j)} \text{ if } g < 1/2 \\ &y = x_{(j)} \text{ if } g = 1/2 \text{ and j is even} \\ &y = x_{(j+1)} \text{ if } g = 1/2 \text{ and j is odd} \\ &y = x_{(j+1)} \text{ if } g > 1/2 \end{aligned}$$

Empirical distribution function:

$$\begin{aligned} &y = x_{(j)} \text{ if } g = 0 \\ &y = x_{(j+1)} \text{ if } g > 0 \end{aligned}$$

Weighted average aimed at $x((N+1)p)$: In that case, we take $(N+1)p = j + g$.

$$y = (1-g)x_{(j)} + gx_{(j+1)}$$

where $x(N+1)$ is replaced by $x(N)$.

Empirical distribution function with averaging:

$$\begin{aligned} &y = \tfrac{1}{2}(x_{(j)} + x_{(j+1)}) \text{ if } g = 0 \\ &y = x_{(j+1)} \text{ if } g > 0 \end{aligned}$$

When weights are associated to the selected variable, the only method available is:

$$y = \begin{cases} x_{(1)} & \text{if } w_{(1)} > pW \\ \frac{1}{2}(x_{(i)} + x_{(i+1)}) & \text{if } \sum_{j=1}^{i} w_{(j)} = pW \\ x_{(i+1)} \text{ if } \sum_{j=1}^{i} w_{(j)} < pW < \sum_{j=1}^{i+1} w_{(j)} \end{cases}$$

where $w(i)$ is the weight associated to $x(i)$ and $W = \sum_{j=1}^{N} w_j$.

Confidence intervals:

You can obtain confidence intervals associated to the quantiles. Two intervals are available:

Confidence interval based on the normal distribution:

The $100 * (1 - alpha)\%$ confidence interval for the p-quantile is:

$$[Np + Z_{\alpha/2}\sqrt{Np(1-p)} + 0.5;\ Np + Z_{1-\alpha/2}\sqrt{Np(1-p)} + 0.5]$$

This kind of interval is valid if the data has a normal distribution and if the sample size is large (>20 observations).

Distribution free confidence interval:

The $100 * (1 - alpha)\%$ confidence interval for the p-quantile is:

$$[x_{(l)}; x_{(u)}]$$

l and u are nearly symmetric around $[Np] + 1$ where $[Np]$ is the integer part of $N \times p$. $x(l)$ and $x(u)$ are the closest to $x([N + 1]p)$ and satisfy:

$$Q(u - 1, n, p) - Q(l - 1, n, p) \geq 1 - \alpha$$

where $Q(k, n, p)$ is the cumulative binomial probability:

$$Q(k, n, p) = \sum_{i=1}^{k} \left(\begin{array}{c} n \\ i \end{array} \right) p^i (1 - p)^{n-i}$$

If weights are selected, confidence intervals cannot be computed.

Charts:

- **Cumulative histogram:** XLSTAT lets you create cumulative histograms by using the empirical cumulative distribution.
- **Box plots**: These univariate representations of quantitative data samples are sometimes called "box and whisker diagrams". It is a simple representation since in the version provided by XLSTAT the 1-st quartile, median and 3-rd quartile are displayed together with both limits (the ends of the "whiskers") beyond which values are considered anomalous. The red line corresponds to the median. Limits are calculated as follows:

Lower limit: $L_{inf} = X(i)$ such that $\{X(i) - [Q1 - 1.5(Q3 - Q1)]\}$ is minimum and $X(i) = Q1 - 1.5(Q3 - Q1)$.

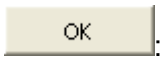
Upper limit: $L_{sup} = X(i)$ such that $\{X(i) - [Q3 + 1.5(Q3 - Q1)]\}$ is minimum and $X(i) = Q3 + 1.5(Q3 - Q1)$

Values that are outside the $]Q1 - 3(Q3 - Q1); Q3 + 3(Q3 - Q1)[$ interval are displayed with the * symbol; values that are in the $[Q1 - 3(Q3 - Q1); Q1 - 1.5(Q3 - Q1)]$ or the $[Q3 + 1.5(Q3 - Q1); Q3 + 3(Q3 - Q1)]$ intervals are displayed with the "o" symbol.

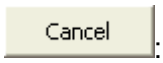
- **Scattergrams**: These univariate representations give an idea of the distribution and possible plurality of the modes of a sample. All points are represented together with the median.

Dialog box

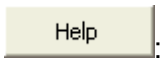
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to samples. If the arrow points to the right, XLSTAT considers that rows correspond to samples and columns to observations.

General tab:

Data: Check this option to select the samples you want to calculate quantiles for.

Estimation method: Choose the method you want to use to calculate the quantiles. A description of the methods can be found in the description section of this help. The default method is weighted average.

Confidence interval:

- **Normal based:** Check this option if you want to display confidence interval based on the normal distribution. See the description section for more details.
- **Distribution free**: Check this option if you want to display distribution free confidence interval. See the description section for more details.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Sample labels: Check this option if the first line of the selections (quantitative data, qualitative date, sub-samples, and weights) contains a label.

Weights: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Sample labels" option is activated.

Sub-sample: Check this option to select a column showing the names or indexes of the sub-samples for each of the observations.

Missing data tab:

Remove observations: Activate this option to ignore an observation that has a missing value.

Estimate missing data: Activate this option to estimate the missing data by using the mean of the sample.

Outputs tab:

Descriptive statistics: Check this option to calculate and display descriptive statistics.

Charts tab:

Empirical cumulative distribution: Activate this option to display the cumulative histograms that actually correspond to the empirical cumulative distribution of the sample.

Show quantile on charts (%): Check this option and enter the percentile to compute the associated value and display it on the charts.

Box plots: Check this option to display box plots (or box-and-whisker plots). See the description section for more details.

Scattergrams: Check this option to display scattergrams. The median (red line) is always displayed.

Results

Summary statistics: This table displays for the selected samples, the number of observations, the number of missing values, the number of non- missing values, the mean and the standard deviation.

Table of quantiles: This table displays percentiles for common values (1, 5, 10, 25, 50, 75, 90, 95, 99) and their associated confidence interval.

Example

An example showing how to compute percentiles is available on XLSTAT Help Center:

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

References

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.

Histograms

Use this tool to create a histogram from a sample of continuous or discrete quantitative data.

In this section:

Description

The histogram is one of the most frequently used display tools as it gives a very quick idea of the distribution of a sample of continuous or discrete data.

Intervals definition

One of the challenges in creating histograms is defining the intervals, as for a determined set of data, the shape of the histogram depends solely on the definition of the classes. Between the two extremes of the single class comprising all the data and giving a single bar and the histogram with one value per class, there are as many possible histograms as there are data partitions.

To obtain a visually and operationally satisfying result, defining classes may require several attempts.

The most traditional method consists of using classes defined by intervals of the same width, the lower bound of the first interval being determined by the minimum value or a value slightly less than the minimum value.

To make it easier to obtain histograms, XLSTAT lets you create histograms either by defining the number of intervals, their width or by specifying the intervals yourself. The intervals are considered as closed for the lower bound and open for the upper bound.

Cumulative histogram

XLSTAT lets you create cumulative histograms either by cumulating the values of the histogram or by using the empirical cumulative distribution. The use of the empirical cumulative distribution is recommended for a comparison with a distribution function of a theoretical distribution.

Comparison to a theoretical distribution

XLSTAT lets you compare the histogram with a theoretical distribution whose parameters have been set by you. However, if you want to check if a sample follows a given distribution, you can use the distribution fitting tool to estimate the parameters of the distribution and if necessary check if the hypothesis is acceptable.

XLSTAT provides the following distributions:

- Arcsine (a): the density function of this distribution (which is a simplified version of the Beta type I distribution) is given by:

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

We have $E(X) = \alpha$ and $V(X) = \alpha(1-\alpha)/2$

- Bernoulli (p): the density function of this distribution is given by:

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

We have $E(X) = p$ and $V(X) = p(1-p)$

The Bernoulli, named after the Swiss mathematician Jacob Bernoulli (1654-1705), allows to describe binary phenomena where only events can occur with respective probabilities of p and $1-p$.

- Beta (a, b): the density function of this distribution (also called Beta type I) is given by:

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

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

We have $E(X) = \alpha/(\alpha+\beta)$ and $V(X) = \alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2]$

- Beta4 (α, β, c, d): the density function of this distribution is given by:

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

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

We have $E(X) = \frac{c+(c-d)\alpha}{(\alpha+\beta)}$ and $V(X) = \frac{(c-d)^2\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$

Pour the type I beta distribution, X takes values in the $[0,1]$ range. The beta4 distribution is obtained by a variable transformation such that the distribution is on a $[c,d]$ interval where c

and d can take any value.

- Binomial (n, p) : the density function of this distribution is given by:

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

We have $E(X) = np$ and $V(X) = np(1-p)$

n is the number of trials, and p the probability of success. The binomial distribution is the distribution of the number of successes for n trials, given that the probability of success is p.

- Negative binomial type I (n, p): the density function of this distribution is given by:

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

We have $E(X) = n(1-p)/p$ and $V(X) = n(1-p)/p^2$

n is the number of successes, and p the probability of success. The negative binomial type I distribution is the distribution of the number x of unsuccessful trials necessary before obtaining n successes.

- Negative binomial type II (k, p): the density function of this distribution is given by:

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

We have $E(X) = kp$ and $V(X) = kp(p+1)$

The negative binomial type II distribution is used to represent discrete and highly heterogeneous phenomena. As k tends to infinity, the negative binomial type II distribution tends towards a Poisson distribution with $\lambda = kp$.

- $Khi^2(df)$: the density function of this distribution is given by:

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

We have $E(X) = df$ and $V(X) = 2df$

The Chi-square distribution corresponds to the distribution of the sum of df squared standard normal distributions. It is often used for testing hypotheses.

- Erlang (k, λ): the density function of this distribution is given by:

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

We have $E(X) = k/\lambda$ and $V(X) = k/\lambda^2$

k is the shape parameter and λ is the rate parameter.

This distribution, developed by the Danish scientist A. K. Erlang (1878-1929) when studying the telephone traffic, is more generally used in the study of queuing problems.

Note: When $k = 1$, this distribution is equivalent to the exponential distribution. The Gamma distribution with two parameters is a generalization of the Erlang distribution to the case where k is a real and not an integer (for the Gamma distribution the scale parameter $\beta = 1/\lambda$ is used).

- Exponential(λ): the density function of this distribution is given by:

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

We have $E(X) = 1/\lambda$ and $V(X) = 1/\lambda^2$

The exponential distribution is often used for studying lifetime in quality control.

- Fisher (df_1, df_2): the density function of this distribution is given by:

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

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

We have $E(X) = df_2/(df_2 - 2)$ if $df_2 > 2$, and $V(X) = \frac{2df_2^2(df_1 + df_2 - 2)}{df_1(df_2-2)^2(df_2-4)}$ if $df_2 > 4$

Fisher's distribution, from the name of the biologist, geneticist and statistician Ronald Aylmer Fisher (1890-1962), corresponds to the ratio of two Chi-square distributions. It is often used for testing hypotheses.

- Fisher-Tippett (β, μ): the density function of this distribution is given by:

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

We have $E(X) = \mu + \beta\gamma$ and $V(X) = (\pi\beta)^2/6$ where γ is the Euler-Mascheroni constant.

The Fisher-Tippett distribution, also called the Log-Weibull or extreme value distribution, is used in the study of extreme phenomena. The Gumbel distribution is a special case of the Fisher-Tippett distribution where $\beta = 1$ and $\mu = 0$.

- Gamma (k, β, μ): the density of this distribution is given by:

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

We have $E(X) = \mu + k\beta$ and $V(X) = k\beta^2$

k is the shape parameter of the distribution and β the scale parameter.

- GEV (β, k, μ): the density function of this distribution is given by:

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

We have $E(X) = \mu + \frac{\beta}{k}\Gamma(1+k)$ and $V(X) = \left(\frac{\beta}{k}\right)^2 \left(\Gamma(1+2k) - \Gamma^2(1+k)\right)$

The GEV (Generalized Extreme Values) distribution is much used in hydrology for modeling flood phenomena. k lies typically between -0.6 and 0.6.

- Gumbel: the density function of this distribution is given by:

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

We have $E(X) = \gamma$ and $V(X) = \pi^2/6$ where γ is the Euler-Mascheroni constant (0.5772156649…).

The Gumbel distribution, named after Emil Julius Gumbel (1891-1966), is a special case of the Fisher-Tippett distribution with $\beta = 1$ and $\mu = 0$. It is used in the study of extreme phenomena such as precipitations, flooding and earthquakes.

- Logistic (μ, s): the density function of this distribution is given by:

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

We have $E(X) = \mu$ and $V(X) = (\pi s)^2/3$

- Lognormal (μ, σ): the density function of this distribution is given by:

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

We have $E(X) = exp(\mu + \sigma^2/2)$ and $V(X) = \left(exp(\sigma^2) - 1\right) exp(2\mu + \sigma^2)$

- Lognormal2 (m, s): the density function of this distribution is given by:

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

We have:

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

And:

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

This distribution is just a reparametrization of the Lognormal distribution.

- Normal (μ, σ) : the density function of this distribution is given by:

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

We have $E(X) = \mu$ and $V(X) = \sigma^2$

- Standard normal: the density function of this distribution is given by:

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

We have $E(X) = 0$ and $V(X) = 1$

This distribution is a special case of the normal distribution with $\mu = 0$ and $\sigma = 1$

- Pareto (a, b): the density function of this distribution is given by:

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

We have $E(X) = ab/(a-1)$ with $V(X) = \frac{ab^2}{((a-1)^2(a-2))}$

The Pareto distribution, named after the Italian economist Vilfredo Pareto (1848-1923), is also known as the Bradford distribution. This distribution was initially used to represent the distribution of wealth in society, with Pareto's principle that 80% of the wealth was owned by 20% of the population.

- PERT (a, m, b): the density function of this distribution is given by:

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

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

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

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

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

The PERT distribution is a special case of the beta4 distribution. It is defined by its definition interval [a, b] and m the most likely value (the mode). PERT is an acronym for *Program Evaluation and Review Technique*, a project management and planning methodology. The PERT methodology and distribution were developed during the project held by the US Navy and Lockheed between 1956 and 1960 to develop the Polaris missiles launched from submarines. The PERT distribution is useful to model the time that is likely to be spent by a team to finish a project. The simpler triangular distribution is similar to the PERT distribution in that it is also defined by an interval and a most likely value.

- Poisson (λ): the density function of this distribution is given by:

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

We have $E(X) = \lambda$ with $V(X) = \lambda$

Poisson's distribution, discovered by the mathematician and astronomer Siméon- Denis Poisson (1781-1840), pupil of Laplace, Lagrange and Legendre, is often used to study queuing phenomena.

- Student (df) : the density function of this distribution is given by:

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

We have $E(X) = 0$ if $df > 1$ with $V(X) = df/(df-2)$ if $df > 2$

The English chemist and statistician William Sealy Gosset (1876-1937), used the nickname Student to publish his work, in order to preserve his anonymity (the Guinness brewery forbade its employees to publish following the publication of confidential information by another researcher). The Student's t distribution is the distribution of the mean of df variables standard normal variables. When $df = 1$, Student's distribution is a Cauchy distribution with the particularity of having neither expectation nor variance.

- Trapezoidal (a, b, c, d): the density function of this distribution is given by:

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

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

This distribution is useful to represent a phenomenon for which we know that it can take values between two extreme values (a and d), but that it is more likely to take values between two values (b and c) within that interval.

- Triangular (a, m, b): the density function of this distribution is given by:

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

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

- TriangularQ (q_1, m, q_2, p_1, p_2): the density function of this distribution is a reparametrization of the Triangular distribution. A first step requires estimating the a and b parameters of the triangular distribution, from the q_1 and q_2 quantiles to which percentages p_1 and p_2 correspond. Once this is done, the distribution functions can be computed using the triangular distribution functions.

- Uniform (a, b): the density function of this distribution is given by:

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

We have $E(X) = (a+b)/2$ with $V(X) = (b-a)^2/12$

The uniform (0,1) distribution is much used for simulations. As the cumulative distribution function of all the distributions is between 0 and 1, a sample taken in a Uniform (0,1) distribution is used to obtain random samples in all the distributions for which the inverse can be calculated.

- Uniform discrete (a, b): the density function of this distribution is given by:

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

We have $E(X) = (a + b)/2$ with $V(X) = [(b - a + 1)^2 - 1]/12$

The uniform discrete distribution corresponds to the case where the uniform distribution is restricted to integers.

- Weibull (β): the density function of this distribution is given by:

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

We have $E(X) = \Gamma(\frac{1}{\beta} + 1)$ with $V(X) = \Gamma(\frac{2}{\beta} + 1) - \Gamma^2(\frac{1}{\beta} + 1)$

β is the shape parameter for the Weibull distribution.

- Weibull (β, γ): the density function of this distribution is given by:

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

We have $E(X) = \gamma\Gamma\left(\frac{1}{\beta} + 1\right)$ with $V(X) = \gamma^2\left[\Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right)\right]$

β is the shape parameter of the distribution and γ the scale parameter. When $\beta = 1$, the Weibull distribution is an exponential distribution with parameter $1/\gamma$.

- Weibull (β, γ, μ): the density function of this distribution is given by:

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

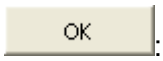
We have $E(X) = \mu + \gamma\Gamma\left(\frac{1}{\beta} + 1\right)$ with $V(X) = \gamma^2\left[\Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right)\right]$

The Weibull distribution, named after the Swede Ernst Hjalmar Waloddi Weibull (1887-1979), is much used in quality control and survival analysis. β is the shape parameter of the distribution and γ the scale parameter. When $\beta = 1$ and $\mu = 0$, the Weibull distribution is an exponential distribution with parameter $1/\gamma$.

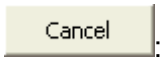
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various

elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to samples. If the arrow points to the right, XLSTAT considers that rows correspond to samples and columns to observations.


: Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Data: Select the quantitative data. If several samples have been selected, XLSTAT will carry out the calculations for each of the samples independently while allowing you to superimpose histograms if you want (see Charts tab). If headers have been selected, check that the "Sample labels" option has been activated.

Data type:

Continuous: Choose this option so that XLSTAT considers your data to be continuous.

Discrete: Choose this option so that XLSTAT considers your data to be discrete.

Subsamples: Activate this option then select a column (column mode) or a row (row mode) containing the sample identifiers. The use of this option gives one histogram per subsample and therefore allows to compare the distribution of data between the subsamples. If a header has been selected, check that the "Sample labels" option has been activated.

- **Variable-Category labels**: Activate this option to use variable-category labels when displaying outputs. Variable-Category labels include the variable name as a prefix and the category name as a suffix.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Sample labels: Activate this option if the first row of the selected data (data, sub-samples, weights) contains a label.

Weights: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Sample labels" option is activated.

Options tab:

Intervals: Choose one of the following options to define the intervals for the histogram:

- **Number**: Choose this option to enter the number of intervals to create.
- **Width**: Choose this option to define a fixed width for the intervals.
- **User defined**: Select a column containing in increasing order the lower bound of the first interval, and the upper bound of all the intervals.
- **Minimum**: Activate this option to enter the value of the lower value of the first interval. This value must be lower or equal to the minimum of the series.

Compare sub-samples: this option is only checked if a column of sub- samples has been selected. Check this option to display the various samples on a single histogram.

- **Compare to total sample**: Check this option so that the descriptive statistics and charts are also displayed for the total sample.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations:

- **For the corresponding sample**: Activate this option to ignore an observation which has a missing value only for samples which have a missing value.
- **For all samples**: Activate this option to ignore an observation which has a missing value for all selected samples.

Estimate missing data: Activate this option to estimate the missing data by using the mean of the sample.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics of the samples.

Charts tab:

Histograms: Activate this option to display the histograms of the samples. For a theoretical distribution, the density function is displayed.

- **Bars**: Choose this option to display the histograms with a bar for each interval.
- **Continuous lines**: Choose this option to display the histograms with a continuous line.

Cumulative histograms: Activate this option to display the cumulated histograms of the samples.

- **Based on the histogram**: Choose this option to display cumulative histograms based on the same interval definition as the histograms.
- **Empirical cumulative distribution**: Choose this option to display cumulative histograms which actually correspond to the empirical cumulative distribution of the sample.

Ordinate of the histograms: Choose the quantity to be used for the histograms: density, frequency or relative frequency.

Display a distribution: Activate this option to compare histograms of samples selected with a density function and/or to compare the histograms of samples selected with a distribution function. Then choose the distribution to be used and enter the values of the parameters if necessary. The **automatic** option allows to let XLSTAT identify the best fitting distribution (determined using a Kolmogorov-Smirnov test).

Results

Summary statistics: This table displays for the selected samples, the number of observations, the number of missing values, the number of non- missing values, the mean and the standard deviation.

Histograms: The histograms are displayed. If desired, you can change the color of the lines, scales, titles as with any Excel chart.

Descriptive statistics for the intervals: This table displays for each interval its lower bound, upper bound, the frequency (number of values of the sample within the interval), the relative frequency (the number of values divided by the total number of values in the sample), and the density (the ratio of the frequency to the size of the interval).

Example

An example showing how to create a histogram is available on XLSTAT Help Center:

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

References

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.

Kernel density estimation

Use this tool to estimate and display the density of a univariate sample using non-parametric methods based on kernels. It is an alternative to the histogram visualization method, and a non-pametric alternative to parametric distribution fitting if you need to re-use the density estimates.

In this section:

Description

Kernel density estimation (KDE) allows to estimate the density of a univariate sample. It is one of the non-parametric alternatives to parametric distribution fitting. While the latter requires knowing the distribution F of the variable, and the knowledge or the estimation of the parameters of F, we will only need here to specify a kernel and a bandwidth.

Let X be a random variable and $\{x_i\}$, $(i = 1, \ldots, n)$, a sample of size n of observations, and w_i the corresponding weights (1s if there is no special weighting). Let $W = \sum_{i=1}^{n} w_i$. The kernel estimate of the probability density function f of X is given by:

$$\hat{f}_h(x) = \frac{1}{Wh} \sum_{i=1}^{n} w_i K\left(\frac{x - x_i}{h}\right)$$

K is the kernel function. h is the bandwidth. Kernel functions are normalized, integrable and symmetric functions. It is also important to note that if K is a kernel then $\lambda K(\lambda z)$ is a kernel as well, which means that one can change the bandwidth h while keeping the mathematical properties unchanged.

While not as much as the bandwidth, the kernel function influences the density. The following kernels are available in XLSTAT:

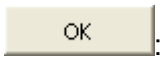
- Biweight (or Quartic): $$ \begin{cases} K(z)=\frac{15}{16}(1-z^2)2 &|z| \leq 1 \ K(z)=0 &|z| \gt 1 \end{cases} $$
- Cosine $$ \begin{cases} K(z)=\frac{1}{2}(1+\cos(z\pi)) &|z| \leq 1 \ K(z)=0 &|z| \gt 1 \end{cases} $$
- Epanechnikov $$ \begin{cases} K(z)=\frac{3}{4}(1- \frac{z^2}{\sqrt{5}})/\sqrt{5} &|z| \leq \sqrt{5} \ K(z)=0 &|z| \gt \sqrt{5} \end{cases} $$
- Epanechnikov (0.75) $$ \begin{cases} K(z)=\frac{3}{4}(1-z^2) &|z| \leq 1 \ K(z)=0 &|z| \gt 1 \end{cases} $$
- Gaussian $$ \begin{cases} K(z)={\frac {1}{\sqrt {2\pi }}}e^{-{\frac {1}{2}}z^{2}} & z \in \left]-\infty, +\infty\right[\ \end{cases} $$

- 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 $$ \begin{cases} K(z)=\frac {4}{3} -8 z^{2+8|z|}3 &|z| \leq 0.5 \ K(z)=\frac {8}{3}(1-|z|)^3 & 0.5 \leq |z| \leq 1 \ K(z)=0 &|z| \gt 1 \end{cases} $$
- Triangular $$ \begin{cases} K(z)=(1-|z|) &|z| \leq 1 \ K(z)=0 &|z| \gt 1 \end{cases} $$
- Tricube $$ \begin{cases} K(z)=\frac{70}{81}(1-|z|^{3)}3 &|z| \leq 1 \ K(z)=0 &|z| \gt 1 \end{cases} $$
- Triweight $$ \begin{cases} K(z)=\frac{35}{32}(1-z^{2)}3 &|z| \leq 1 \ K(z)=0 &|z| \gt 1 \end{cases} $$
- Uniform $$ \begin{cases} K(z)=0.5 &|z| \leq 0.5 \ K(z)=0 &|z| \gt 0.5 \end{cases} $$

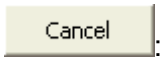
As for the number of bins of histograms, the bandwidth h influences a lot the shape of the density function, including the number of modes. XLSTAT offers the following options for the bandwidth: * User defined: You can enter the value of your choice * Silverman(1): $h = 0.9 * min(\hat{\sigma}, IQR/1.34)/n^{1/5}$, where IQR is the unscaled interquartile range (see Silverman, page 48) * Silverman(2): $h = 1.06 * \hat{\sigma}/n^{1/5}$, (see Silverman, page 45)

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.


: Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Data: Select the quantitative data. If several samples have been selected, XLSTAT will carry out the calculations for each of the samples independently. If headers have been selected, check that the "Sample labels" option has been activated.

Weights: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Sample labels" option is activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Sample labels: Activate this option if the first row of the selected data (data, sub-samples, weights) contains a label.

Kernel: the kernel function that will be used (see the description section).

Bandwidth: XLSTAT allows you to choose a method for automatically computing the bandwidths (see the description section):

- **User defined**: the bandwidth is constant and equal to the fixed value. Enter the value of the bandwidth.
- **Silverman (1)**: This is the default option.
- **Silverman (2)**.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations:

- **For the corresponding sample**: Activate this option to ignore an observation which has a missing value only for samples which have a missing value.
- **For all samples**: Activate this option to ignore an observation which has a missing value for all selected samples.

Estimate missing data: Activate this option to estimate the missing data by using the mean of the sample.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics of the samples.

Kernel density estimates: Activate this option to display the kernel density estimates for each observation.

Charts tab:

Densité de noyau : activez cette option pour afficher la courbe de densité de noyau. Utilisez les options suivantes pour adapter l'affichage:

- **Number of points**: Enter the number of points at which the density is computed. If the minimum and the maximum are not specified (see below), the computations are done within the interval $]min - 0.1 \times range, max + 0.1 \times range[$.
- **Minimum**: Activate this option to enter the lower value at which the density curve is computed.
- **Maximum**: Activate this option to enter the upper value at which the density curve is computed.

Histograms: Activate this option to display a histogram behind the kernel density curve. Use the following options to control the display of the histogram:

- **Intervals**: Choose one of the following options to define the intervals for the histogram:
- **Number**: Choose this option to enter the number of intervals to create.
- **Width**: Choose this option to define a fixed width for the intervals.
- **User defined**: Select a column containing in increasing order the lower bound of the first interval, and the upper bound of all the intervals.
- **Minimum**: Activate this option to enter the value of the lower value of the first interval. This value must be lower or equal to the minimum of the series.

Results

Summary statistics: This table displays for the selected samples, the number of observations, the number of missing values, the number of non- missing values, the mean and the standard deviation.

The following results are displayed for each selected sample:

Bandwith: XLSTAT displays the bandwidth that was used for the computations. If you want to modify the kernel density plot, you might want to change the value of the bandwidth based on the value provided here: increase/decrease the value of the bandwidth if you want a smoother/more detailed curve.

Kernel density estimates: This table displays the kernel density estimates for each observation.

Charts: XLSTAT displays the kernel density curves. If requested, histograms are displayed behind the density curves. If desired, you can change the color of the lines, scales, titles as with any Excel chart.

Descriptive statistics for the intervals: This table displays for each interval of the histogram, its lower bound, upper bound, the frequency (number of values of the sample within the interval), the relative frequency (the number of values divided by the total number of values in the sample), and the density (the ratio of the frequency to the size of the interval).

Example

An example of Kernel Density Estimation is available on the XLSTAT Help Center. To consult the tutorial, please go to:

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

References

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.

Normality tests

Use this tool to check if a sample can be considered to follow a normal distribution. The distribution fitting tool enables the parameters of the normal distribution to be estimated but the tests offered are not as suitable as those given here.

In this section:

Description

Assuming a sample is normally distributed is common in statistics. But checking that this is actually true is often neglected. For example, the normality of residuals obtained in linear regression is rarely tested, even though it governs the quality of the confidence intervals surrounding parameters and predictions.

XLSTAT offers four tests for testing the normality of a sample:

The Shapiro-Wilk test which is best suited to samples of less than 5000 observations;

The Anderson-Darling test proposed by Stephens (1974) is a modification of the Kolmogorov-Smirnov test and is suited to several distributions including the normal distribution for cases where the parameters of the distribution are not known and have to be estimated;

The Lilliefors test is a modification of the Kolmogorov-Smirnov test and is suited to normal cases where the parameters of the distribution, the mean and the variance are not known and have to be estimated;

The Jarque-Bera test which is more powerful the higher the number of values.

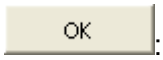
In order to check visually if a sample follows a normal distribution, it is possible to use P-P plots and Q-Q plots:

P-P Plots (normal distribution): P-P plots (for Probability-Probability) are used to compare the empirical distribution function of a sample with that of a sample distributed according to a normal distribution of the same mean and variance. If the sample follows a normal distribution, the points will lie along the first bisector of the plan.

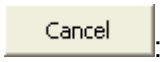
Q-Q Plots (normal distribution): Q-Q plots (for Quantile-Quantile) are used to compare the quantities of the sample with those of a sample distributed according to a normal distribution of the same mean and variance. If the sample follows a normal distribution, the points will lie along the first bisector of the plan.

Dialog box

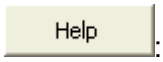
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to samples. If the arrow points to the right, XLSTAT considers that rows correspond to samples and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Data: Select the quantitative data. If several samples have been selected, XLSTAT carries out normality tests for each of the samples independently. If headers have been selected, check that the "Sample labels" option has been activated.

Weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Sample labels" option is activated.

Shapiro-Wilk test: Activate this option to perform a Shapiro-Wilk test.

Anderson-Darling test: Activate this option to perform an Anderson- Darling test.

Lilliefors test: Activate this option to carry out a Lilliefors test.

Jarque-Bera test: Activate this option to carry out a Jarque-Bera test.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Sample labels: Activate this option if the first row of the selected data (data, sub-samples, weights) contains a label.

Significance level (%): Enter the significance level for the tests.

Subsamples: Activate this option then select a column (column mode) or a row (row mode) containing the sample identifiers. The use of this option gives one series of tests per subsample. If a header has been selected, check that the "Sample labels" option has been activated.

Missing data tab:

Remove observations:

- **For the corresponding sample**: Activate this option to ignore an observation which has a missing value only for samples which have a missing value.
- **For all samples**: Activate this option to ignore an observation which has a missing value for all selected samples.

Estimate missing data: Activate this option to estimate the missing data by using the mean of the sample.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics of the samples.

Charts tab:

P-P plots: Activate this option to display Probability-Probability plots based on the normal distribution.

Q-Q Plots: Activate this option to display Quantile-Quantile plots based on the normal distribution.

Results

For each test requested, the statistics relating to the test are displayed including, in particular, the p-value which is afterwards used in interpreting the test by comparing with the chosen significance threshold.

If requested, P-P and Q-Q plots are then displayed.

Example

An example showing how to test the normality of a sample is available on the XLSTAT Help Center:

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

References

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.

Resampling

Use this tool to calculate descriptive statistics using resampling methods (bootstrap, jackknife…) for a set of quantitative variables.

In this section:

Description

Resampling methods have become more and more popular since computational power has increased. It is a well-known approach to nonparametric statistics. The principle is very simple: from your original sample, randomly draw a new sample and recalculate statistics. Repeating this step many times gives you the empirical distribution of the statistic, from which you obtain the standard error, and confidence intervals.

With XLSTAT, you can apply these methods on a selected number of descriptive statistics for quantitative data.

Three resampling methods are available:

- **Bootstrap**: It is the most famous approach; it has been introduced by Efron and Tibisharni (1993). It is a statistical method for estimating the sampling distribution of an estimator by sampling with replacement from the original sample. The number of samples has to be given.
- **Random without replacement**: Subsamples are drawn randomly from the original sample. The size of the subsample has to be specified.
- **Jackknife**: The sampling procedure is based on suppressing one observation to the original sample (of size n). Each subsample has $n-1$ observations and the process is repeated n times. It is less robust than the bootstrap.

Although you can select several variables (or samples) at the same time, XLSTAT calculates all the descriptive statistics for each of the samples independently.

Descriptive statistics for quantitative data:

Let's consider a sample made up of n items of quantitative data $\{x_1, x_2, \ldots, x_n\}$ whose respective weights are $\{w_1, w_2, \ldots, w_n\}$.

- **Sum** *: The weighted sum of the values is defined by:

$$S = \sum_{i=1}^{n} w_i x_i$$

- **Mean** *: The mean of the sample is defined by:

$$\mu = \frac{S}{S_W}$$

- **Variance (n)** *: The variance of the sample is defined by:

$$s(n)^2 = \frac{\sum_{i=1}^{n} w_i (x_i - \mu)^2}{S_W}$$

Note 1: When all the weights are 1, the variance is the sum of the square deviation to the mean divided by n, hence its name.

Note 2: The variance (n) is a biased estimate of the variance which assumes that the sample is a good representation of the total population. The variance $(n-1)$ is, on the other hand, calculated taking into account an approximation associated with the sampling.

- **Variance (n-1)** *: The estimated variance of the sample is defined by:

$$s(n-1)^2 = \frac{\sum_{i=1}^{n} w_i (x_i - \mu)^2}{S_W - S_W/n}$$

Note 1: When all the weights are 1, the variance is the sum of the square deviation to the mean divided by $n-1$, hence its name.

Note 2: The variance (n) is a biased estimate of the variance which assumes that the sample is a good representation of the total population. The variance $(n-1)$ is, on the other hand, calculated taking into account an approximation associated with the sampling.

- **Standard deviation (n)** *: The standard deviation of the sample is defined by: $s(n)$
- **Standard deviation (n-1)** *: The standard deviation of the sample is defined by: $s(n-1)$
- **Median** *: The median $Q2$ is the value for which 50% of the values are less than $Q2$.
- **1st quartile** *: The first quartile $Q1$ is the value for which 25% of the values are less than $Q1$

- **3rd quartile** *: The third quartile $Q3$ is the value for which 75% of the values are less than $Q3$

- **Variation coefficient** *: this coefficient is only calculated if the mean of the sample is non-zero. It is defined by: $CV = s(n)/\mu$. This coefficient measures the dispersion of a sample relative to its mean. It is used to compare the dispersion of samples whose scales or means differ greatly.

- **Standard error of the mean** *: this statistic is defined by:

$$s_\mu = \frac{s_n}{\sqrt{n-1}}$$

- **Mean absolute deviation** *: as for standard deviation or variance, this coefficient measures the dispersion (or variability) of the sample. It is defined by:

$$e(\mu) = \frac{\sum_{i=1}^{n} w_i |x_i - \mu|}{S_W}$$

- **Median absolute deviation** *: this statistic is the median of absolute deviations to the median.

- **Geometric mean** *: this statistic is only calculated if all the values are strictly positive. It is defined by:

$$\mu_G = exp\left(\frac{1}{S_W}\sum_{i=1}^{n} w_i Ln(x_i)\right)$$

If all the weights are equal to 1, we have:

$$\mu_G = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

- **Geometric standard deviation** *: this statistic is defined by:

$$\sigma_G = \exp\left(\frac{1}{S_W}\sum_{i=1}^{n} w_i \left(\operatorname{Ln}\left(x_i\right) - \operatorname{Ln}\left(\mu_G\right)\right)^2\right)$$

- **Harmonic mean** *: this statistic is defined by:

$$\mu_H = \frac{S_W}{\sum_{i=1}^{n} \frac{w_i}{x_i}}$$

Statistics followed by an asterisk (*) take the weights of observations into account.

Statistics obtained after resampling:

Let S be one of the preceding statistics, during the resample procedure it has been computed B times. In the case of bootstrap and random without replacement, we have:

Mean: It is the mean on the B samples:

$$\hat{\mu}^{\star}(S) = \frac{\sum_{i=1}^{B} \hat{S}_i}{B}$$

where S_i is the estimated value of S for sample i.

Standard error:

$$\hat{\sigma}^{*}(S) = \sqrt{\frac{\sum_{i=1}^{B} \left(\hat{S}_i - \hat{\mu}^{*}(S)\right)^2}{B-1}}$$

Standard bootstrap confidence interval: It is defined by:

$$[S \pm u_{1-\alpha/2}\hat{\sigma}^{*}(S)]$$

where u is the $1-\alpha/2$ percentile of the normal distribution and $1-\alpha$ is the confidence degree. This type of interval depends on a parametric distribution.

Simple percentile confidence interval: Confidence interval limits are obtained using the $\alpha/2$ and $1-\alpha/2$ percentiles of the empirical distribution of S.

Bias corrected percentile confidence interval: Confidence interval limits are also obtained using percentiles of the empirical distribution of S, but with a small difference. These limits are noted S_{a_1} and S_{a_2}. Let p be the proportion of S_i lower than S (value of the statistic on the original sample). U_p is the percentile associated to the normal distribution with probability p. Then we have: $a_1 = \Phi(2u_p + u_{\alpha/2})$ and $a_2 = \Phi(2u_p + u_{1-\alpha/2})$. For more details on this approach please refer, to Efron and Tibshirani (1993).

Jackknife :

- **Mean**:

$$\hat{\mu}(S) = \frac{\sum_{i=1}^{n} \hat{S}_{(-i)}}{n}$$

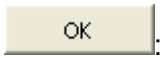
where $S(-i)$ is obtained on the sample without observation i.

- **Standard error**:

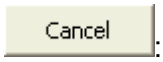
$$\hat{\sigma}^*(S) = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{S}_{(-i)} - \hat{\mu}^*(S))^2}$$

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to samples. If the arrow points to the right, XLSTAT considers that rows correspond to samples and columns to observations.

General tab:

Quantitative data: Select the samples of quantitative data you want to calculate descriptive statistics for.

Method: Choose the resampling method you want to use.

- **Bootstrap**: Check this button to apply the bootstrap method.
- **Random without replacement**: Check this button to apply the random without replacement method.
- **Jackknife**: Check this button to apply the Jackknife approach.

Sample size: Enter the size of the subsample in the case of random without replacement.

Number of sample: Enter the number of sample in the case of the bootstrap and the random without replacement.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Sample labels: Check this option if the first line of the selections (quantitative data, qualitative date, sub-samples, and weights) contains a label.

Weights: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Sample labels" option is activated.

Missing data tab:

Remove observations: Activate this option to ignore an observation that has a missing value.

Estimate missing data: Activate this option to estimate the missing data by using the mean of the sample.

Outputs tab:

Quantitative Data: Activate the options for the descriptive statistics you want to calculate. The various statistics are described in the description section.

- **All**: Click this button to select all.
- **None**: Click this button to deselect all.
- **Display vertically**: Check this option so that the table of descriptive statistics is displayed vertically (one line per descriptive statistic).

Confidence interval: Enter the size of the confidence interval (in %).

Standard bootstrap confidence interval: Activate this option to display the standard bootstrap confidence interval.

Simple percentile confidence interval: Activate this option to display the simple percentile confidence interval.

Bias corrected percentile confidence interval: Activate this option to display the bias corrected percentile confidence interval.

Resampled statistics: Activate this option to display the resampled statistics.

Resampled data: Activate this option to display the resampled data.

Charts tab:

Histograms: Activate this option to display the histograms of the samples. For a theoretical distribution, the density function is displayed.

- **Bars**: Choose this option to display the histograms with a bar for each interval.
- **Continuous lines**: Choose this option to display the histograms with a continuous line.

Cumulative histograms: Activate this option to display the cumulative histograms of the samples.

- **Based on the histogram**: Choose this option to display cumulative histograms based on the same interval definition as the histograms.
- **Empirical cumulative distribution**: Choose this option to display cumulative histograms that actually correspond to the empirical cumulative distribution of the sample.

Ordinate of the histograms: Choose the quantity to be used for the histograms: density, frequency or relative frequency.

Results

Summary statistics: This table displays for the selected samples, the number of observations, the number of missing values, the number of non- missing values, the mean and the standard deviation.

Resampling: This table displays for the selected statistics, the mean, the standard error and the confidence interval obtained with resampling.

Resampled statistics: This table displays the resampled statistics for each of the B samples.

Resampled data: This table displays the B samples obtained by resampling the initial data.

Histograms: The histograms are displayed. If desired, you can change the color of the lines, scales, titles as with any Excel chart.

Descriptive statistics for the intervals: This table displays for each interval its lower bound, upper bound, the frequency (number of values of the sample within the interval), the relative frequency (the number of values divided by the total number of values in the sample), and the density (the ratio of the frequency to the size of the interval).

Example

An example showing how to create apply bootstrap is available on XLSTAT Help Center:

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

References

Efron B. and Tibshirani R.J. (19 93). An introduction to the bootstrap, Chapman & Hall / CRC.

Good P. (2006). Resampling methods. A guide to data analysis. Third Edition. Birkhäuser.

Similarity/dissimilarity matrices (Correlations, ...)

Use this tool to calculate a proximity index between the rows or the columns of a data table. The most classic example of the use of this tool is in calculating a correlation or covariance matrix between quantitative variables.

In this section:

Description

This tool offers a large number of proximity measurements between a series of objects whether they are in rows (usually the observations) or in columns (usually the variables).

The correlation coefficient is a measurement of the similarity of the variables: the more similar the variables are, the higher the correlation coefficient.

Similarities and dissimilarities

The proximity between two objects is measured by measuring to what extent they are similar (similarity) or dissimilar (dissimilarity).

The available simiilarity and dissimilarity indexes depend on the nature of the data:

- Quantitative data:

The **similarity** coefficients proposed by the calculations from the quantitative data are as follows: Cosine, Covariance (n-1), Covariance (n), Inertia, Gower coefficient, Kendall correlation coefficient, Partial correlation coefficient, Pearson correlation coefficient, Percent agreement, Spearman correlation coefficient.

The **dissimilarity** coefficients proposed by the calculations from the quantitative data are as follows: Bhattacharya's distance, Bray and Curtis' distance, Canberra's distance, Chebychev's distance, Chi² distance, Chi² metric, Chord distance, Squared chord distance, Euclidian distance, Squared Euclidian distance, Geodesic distance, Kendall's dissimilarity, Mahalanobis distance, Manhattan distance, Pearson's dissimilarity, Percent disagreement, Spearman's dissimilarity.

- Binary data:

The **similarity** and **dissimilarity** (pay simple transformation) coefficients proposed by the calculations from the binary data are as follows: Co-occurrence, Dice coefficient (also known as the Sorensen coefficient), Jaccard coefficient (1), Jaccard coefficient (2), Rand coefficient, Adjusted Rand coefficient, Kulczinski coefficient, Pearson Phi, Percent agreement, Ochiai coefficient, Rogers & Tanimoto coefficient, Sokal & Michener's coefficient (simple matching coefficient), Sokal & Sneath's coefficient (1), Sokal & Sneath's coefficient (2).

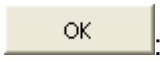
- Qualitative data:

The **similarity** coefficients proposed by the calculations from the qualitative data are as follows: Cooccurrence, Percent agreement.

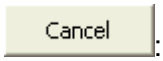
The **dissimilarity** indexes proposed by the calculations from the qualitative data are as follows: Percent disagreement

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find the description of the various elements of the dialog box below.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Data: Select a table comprising N objects described by P descriptors. If column headers have been selected, check that the "Column labels" option has been activated.

Data type: Choose the type of data selected.

Note : in the case where the selected data type is « Qualitative », whatever the true type of the data, they are considered as qualitative.

Row weights: Activate this option if the rows are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column labels" option is activated.

Proximity type: **similarities / dissimilarities**: Choose the proximity type to be used. The data type and proximity type determine the list of possible indexes for calculating the proximity matrix.

Note: to calculate a classical correlation coefficient (also called Pearson's correlation coefficient) you must select data types "quantitative", "similarities" and "Pearson's correlation coefficient".

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selections (Observations/variables table, row labels, row weights, column weights) contains a label.

Row labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Column labels" option is activated you need to include a header in the selection. If this option is not activated, the observation labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Compute proximities for:

Columns: Activate this option to measure proximities between columns.

Rows: Activate this option to measure proximities between rows.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Pairwise deletion: Activate this option to remove observations with missing data only when the variables involved in the calculations have missing data. For example, when calculating the correlation between two variables, an observation will be ignored only if the data corresponding to one of the two variables is missing.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

Flag similar objects: Activate this option to identify similar objects in the proximity matrix.

List similar objects: Activate this option to display the list of similar objects.

Dissimilarity threshold: Enter the threshold value of the index from which you consider objects to be similar. If the chosen index is a similarity, the values will be considered as being similar if they are greater than this value. If the chosen index is a dissimilarity, the values will be considered as being similar if they are less than this value.

Cronbach's Alpha: Activate this option to calculate Cronbach's alpha coefficient.

Bartlett's sphericity test: Activate this option to calculate Bartlett's sphericity test (only for Pearson correlation or covariance).

Significance level (%): Enter the significance level for the sphericity test.

Results

Summary statistics: This table shows the descriptive statistics for the samples.

Proximity matrix: This table displays the proximities between the object for the chosen index. If the "Identify similar objects" option has been activated and the dissimilarity threshold has been exceeded, the values for the similar objects are displayed in bold.

List of similar objects: If the "List similar objects" option has been checked and at least one pair of objects has a similarity beyond the threshold, the list of similar objects is displayed.

Example

An example showing how to compute a dissimilarity matrix is available on the XLSTAT Help Center:

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

References

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.

Biserial correlation

Use this tool to compute the biserial correlation between on one hand, one or more quantitative variables, and on the other hand, one or more qualitative binary variables.

In this section:

Description

This tool allows computing the biserial correlation between on one hand, one or more quantitative variables, and on the other one or more binary variables. The biserial correlation introduced by Pearson (1909), between a quantitative variable and a binary variable is given by:

$$r = \frac{(\hat{\mu}_2 - \hat{\mu}_1)}{\hat{\sigma}_n}\sqrt{p_1 p_2}$$

Where $\hat{\mu}_1$ and $\hat{\mu}_2$ are the estimated means for the two possible values of the binary variable, $\hat{\sigma}_n$ is the biased standard deviation estimated on all the data, and p_1 and p_2 are the proportions corresponding to the two values of the binary variable ($p_1 + p_2 = 1$). As for the Pearson correlation, the biserial correlation coefficient varies between -1 and 1. 0 corresponds to no association (the means of the quantitative variable for the two categories of the qualitative variable are identical).

XLSTAT allows testing if the r value that has been obtained is different from 0 or not.

For the two-tailed test, the null H0 and alternative Ha hypotheses are as follows:

- H0: $r = 0$
- Ha: $r \neq 0$

In the left one-tailed test, the following hypotheses are used:

- H0: $r = 0$
- Ha: $r < 0$

In the right one-tailed test, the following hypotheses are used:

- H0: $r = 0$
- Ha: $r > 0$

Two methods to compute the p-value are proposed by XLSTAT. The user can choose between a p-value computed using on a large sample approximation, and a p-value computed using Monte Carlo resamplings. The second method is recommended.

To compute the p-value using the large sample approximation, we use the following result:

If n is the full sample size, the statistic defined by

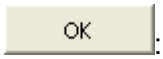
$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

follows a Student distribution with $n - 2$ degrees of freedom under the null hypothesis.

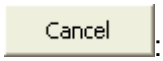
Note: the XLSTAT_Biserial spreadsheet function can be used to compute the biserial correlation between a quantitative variable and a binary variable.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Qualitative variables: Activate this option to select one or more quantitative variables. If a column header has been selected, check that the "Variable labels" option is activated.

Qualitative variables: Activate this option to select one or more binary qualitative variables. If a column header has been selected, check that the "Variable labels" option is activated.

Control category : choose the category you want to assign as the group 2 for the above calculations.

Weights: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (variables, weights, observations labels) includes a header.

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Significance level (%): Enter the significance level for the test (default value: 5%).

Asymptotic p-value: Activate this option if you want XLSTAT to calculate the p-value based on the asymptotic approximation (see description).

Monte Carlo method: Activate this option if you want XLSTAT to calculate the p-value based on Monte Carlo permutations, and select the number of random permutations to perform or the maximum time to spend.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

- **For the corresponding variable**: Activate this option to ignore an observation which has a missing value only for the variables that have a missing value.

- **For all variables**: Activate this option to ignore an observation which has a missing value for all selected variables.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables. For qualitative variables, the categories with their respective frequencies and percentages are displayed.

The biserial correlation is then given for each pair (quantitative variable, qualitative variable). The p-values are then displayed if they have been requested. The details for the test are given only when the correlation is calculated one quantitative variable and one qualitative variable.

Example

An example showing how to compute the biserial correlation is available on the XLSTAT Help Center. To download this data, go to:

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

References

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.

Multicolinearity statistics

Use this tool to identify multicolinearities between your variables.

In this section:

Description

Variables are said to be multicolinear if there is a linear relationship between them. This is an extension of the simple case of colinearity between two variables. For example, for three variables X_1, X_2 and X_3, we say that they are multicolinear if we can write:

$$X_1 = aX_2 + bX_3$$

where a and b are real numbers.

Principle Component Analysis (PCA) can detect the presence of multicolinearities within the data (a number of non-null factors less than the number of variables indicates the presence of a multicolinearity), but it cannot identify the variables which are responsible.

To detect the multicolinearities and identify the variables involved, linear regressions must be carried out on each of the variables as a function of the others. We then calculate:

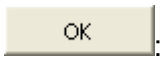
- The R^2 of each of the models. If the R^2 is 1, then there is a linear relationship between the dependent variable of the model (the Y) and the explanatory variables (the X).
- The **tolerance** for each of the models. The tolerance is ($1-R^2$). It is used in several methods (linear regression, logistic regression, discriminant factorial analysis) as a criterion for filtering variables. If a variable has a tolerance less than a fixed threshold (the tolerance is calculated by taking into account variables already used in the model), it is not allowed to enter the model as its contribution is negligible and it risks causing numerical problems.
- The VIF (Variance Inflation Factor) which is equal to the inverse of the tolerance.

Detect multicolinearities within a group of variables can be useful especially in the following cases:

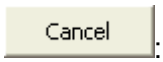
- To identify structures within the data and take operational decisions (for example, stop the measurement of a variable on a production line as it is strongly linked to others which are already being measured),
- To avoid numerical problems during certain calculations. Certain methods use matrix inversions. The inverse of a $(p \times p)$ matrix can be calculated if it is of rank p (or regular). If it is of lower rank, in other words, if there are linear relationships between its columns, then it is singular and cannot be inverted.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Observations/variables table: Select a table with N observations and P variables. If column headers have been selected, check that the "Variable labels" option has been activated.

Variable labels: Activate this option if the first row of the selection includes a header.

Weights: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

Correlations: Activate this option to display the correlations matrix.

R²: Activate this option to display the R-squares.

Tolerance: Activate this option to display the tolerances.

VIF: Activate this option to display the VIFs.

Charts tab:

Bar charts: Activate this option to display the bar charts of the following statistics:

- R²
- Tolerance
- VIF

Results

The results comprise the descriptive statistics of the variables selected, the correlation matrix of the variables and the multicolinearity statistics (R², Tolerance and VIF). Bar charts are used to locate the variables which are more multi-correlated than the others.

When the tolerance is 0, the VIF has infinite value and is not displayed.

Example

References

Belsley D.A., Kuh E. and Welsch R.E. (1980). Regression Diagnostics, Identifying Influential Data and Sources of Collinearity. Wiley, New York.

Reliability Analysis

Reliability analysis can be used to characterize measurement scales composed of elements (questions in the case of a questionnaire). The procedure computes several measures to evaluate the reliability of a scale and also provides information on relationships between its elements.

In this section:

Description

Reliability analysis is used in several areas, noticeably in social sciences. The terminology finds its origin in psychometry. We define a test made up of questions. Questions are called elements. Elements are gathered within homogeneous constructs called factors, measurement scales, latent variables or concepts.

As an illustration, graphical skill might be a factor on which we wish to measure a level on a scale of measurement. The goal of reliability analysis is to assess the reliability of this scale of measurement, or, in other words, that the construct questions are coherent and measure the same thing. In the case of the graphical skill, a question on mental calculus would downgrade the coherence of the scale.

To the statistician, questions are variables often measured on Likert-type scale (rating answers). Results of a test collected on a group of individuals are gathered in an individuals/variables array. To ensure compatibility with other areas such as quality control where the reliability analysis might also be used, those arrays are labelled observations/variables within XLSTAT.

Methods implemented in XLSTAT are used to estimate the internal consistency of a scale by making sure that results to different questions addressing the same phenomenon are coherent. Moreover, they also measure the reliability between two tests administered to the same individuals at two different times.

"Internal" analysis allows you to determine which elements of a survey might be correlated by providing an index related to the internal consistency of the scale. It also allows you to identify unnecessary elements that could be removed from the scale.

The "split-half reliability" analysis measures the equivalence between two parts of a test (parallel forms reliability). This type of analysis is used for two similar sets of items measuring the same thing, using the same instrument and with the same people.

Cronbach’s alpha

Cronbach's alpha index measures internal consistency, which is, how closely related a set of items are. It is considered to be a measure of scale reliability.

This index (represented by the Greek letter « α ») is the mathematical equivalent of the average of all correlations between 2 equal portions of the scale.

The formula of the raw Cronbach’s alpha is given by:

$$\alpha = \frac{\sum_{h \neq h'} \mathrm{cov}(x_h, x_{h'})}{\mathrm{var}\left(\sum_h x_h\right)} \times \frac{p}{p-1}$$

With x_h the h^{th} item of the scale and p the total number of items.

Many methodologists recommend a minimum alpha coefficient between 0.65 and 0.8 (or more); those below 0.5 are usually considered as unacceptable.

XLSTAT also provides the standardized Alpha coefficient which is equivalent to the reliability that would be obtained if all items values were standardized (centered-reduced variables) before computations.

The formula of the standardized Cronbach’s alpha is given by the following equation:

$$\alpha_{std} = \frac{\sum_{h \neq h'} \mathrm{cor}(x_h, x_{h'})}{p + \sum_{h \neq h'} \mathrm{cor}(x_h, x_{h'})} \times \frac{p}{p-1}$$

Guttman's reliability coefficients (lambda 1-6)

There are six reliability coefficients: L1 to L6. Four of them (L1, L3, L5 and L6) are used to estimate internal consistency and the last two (L2 and L4) are used for split-half reliabilities:

L1: Intermediate coefficient used in computing the other lambdas.

L2: Estimation of the inter-score correlation in the case of parallel measurements. It is more complex than the Cronbach alpha and better represents the true reliability of the test.

L3: Equivalent to Cronbach's alpha.

L4: Guttman split-half reliability (See the description below).

L5: Recommended when a single item strongly covaries with other items whereas those items don’t covary between each other.

L6: Recommended when inter-item correlations are low compared to the item squared multiple correlations (becomes a better estimator as the number of elements increase).

Spearman-Brown reliability (Split-half model):

Another way to calculate the reliability of a scale is to randomly split it into two parts. If the scale is perfectly reliable, we expect the two halves to be perfectly correlated (ie, $R = 1$). Bad reliability leads to imperfect correlations. We can estimate that reliability with the Spearman-Brown coefficient formula:

$$Y = \frac{2R}{1 + R}$$

Where Y is the Spearman-Brown split-half coefficient and R represents the correlation between the two halves of the scale.

When the two halves have different sizes, a more accurate estimate of the reliability is used (Horst's Formula) which reads:

$$H = \frac{-R^2 + \sqrt{R^4 + 4R^2(1 - R^2)k_1k_2/k}}{2(1 - R^2)k_1k_2/k}$$

Where H is the Horst split-half coefficient, R represents the correlation between the two halves of the scale, k_1 the number of items in the first part, k_2 the number of items in the second part and k the total number of items in the scale.

Guttman split-half reliability (Split-half model):

It is similar to the Spearman-Brown half-reliability coefficient, but does not consider the reliabilities or variances to be equal in both halves (Tau- equivalence). It is calculated as follows:

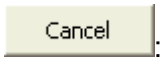
$$G = \frac{2(S_p^2 - S_{p_1}^2 - S_{p_2}^2)}{S_p^2}$$

Where G is the Guttman split-half reliability coefficient (L4),S_p,S_{p_1}, S_{p_2} respectively the variances of total and subparts of the test.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Observations/items table: Select a table made up of N observations described by P items. If column headers have been selected, check that the "Variable labels" option has been activated.

Observations/items table (1): Select a table made up of $N1$ observations described by P items (first split-half). If column headers have been selected, check that the "Variable labels" option has been activated.

Observations/items table (2): Select a table made up of $N2$ observations described by P items (second split-half). If column headers have been selected, check that the "Variable labels" option has been activated.

Weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Type of reliability: Choose the type of reliability to use for the computations (see the description section for more details).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (row and column variables, weights) includes a header.

Options tab:

Enumerate: Activate this option to find the optimal split-half maximizing the Guttman L4 reliability index for the given input scale. It is carried out by testing every combination in two

parts of the initial test.

- **Maximum time (s)**: Activate this option to set a maximum delay in seconds in order to search for the optimal Guttman L4 index (max value).
- **Fast** : Use a maximum search algorithm providing an optimal solution in a faster time (preferable when the number of items becomes larger)

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Pairwise deletion: Activate this option to remove observations with missing data only when the variables involved in the calculations have missing data. For example, when calculating the correlation between two variables, an observation will only be ignored if the data corresponding to one of the two variables is missing.

Estimate missing data: Activate this option to estimate the missing data before the calculation starts.

- **Mean or mode**: Activate this option to estimate the missing data by using the mean (quantitative variables) or the mode (qualitative variables) for the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data for an observation by searching for the nearest neighbor to the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

Deleted items statistics: Activate this option to calculate and display statistics on the comparison of each element to the scale composed of others. These statistics include the mean and variance of the scale if the element was deleted, the correlation between the element and the sum of others, Cronbach's alpha if the element was removed from the scale, Guttman's L6 if the element was removed from the scale and the coefficient of determination (R^2) between the deleted element and the scale composed of others.

Correlation matrix: Activate this option to display the correlation matrix that corresponds to the Pearson correlation applied to the observations/items table given as input.

Covariance matrix: Activate this option to display the covariance matrix related to the observations/items table given as input.

Cronbach's alpha statistics: Activate this option to display the raw and standardized Cronbach's alphas for the global scale and subscales (only when the "split-half model" reliability is selected).

Guttman's statistics: Activate this option to display the Guttman's reliability indices.

- **Display the best split-half**: Activate this option to display each partition corresponding to an optimal Guttman L4 coefficient.

Split-half model's statistics: Activate this option to display the reliability indices corresponding to a split-half reliability model (correlation between the two halves of the scale, Spearman-Brown coefficient or Horst coefficient, Guttman split-half reliability coefficient (L4) for the partitions given as input)

Charts tab:

Correlation maps: Several visualizations of a correlation matrix are proposed.

- The "**blue-red** " option allows to represent low correlations with cold colors (blue is used for the correlations that are close to -1) and the high correlations are with hot colors (correlations close to 1 are displayed in red color).
- The "**Black and white** " option allows to either display in black the positive correlations or in white the negative correlations (the diagonal of 1s is display in grey color), or to display in black the significant correlations, and in white the correlations that are not significantly different from 0.
- The "**Patterns** " option allows to represent positive correlations by lines that rise from left to right, and the negative correlations by lines that rise from right to left. The higher the absolute value of the correlation, the large the space between the lines.

Results

The correlation matrix and descriptive statistics of scale / item are displayed.

Cronbach's alpha / Guttman's statistics estimate the reliability of the scale entered as input, while the deleted items statistics provide useful information on the influence of the withdrawal of each element on the overall reliability of the scale.

The internal reliability model allows the computation of the best partition which can be used in the case where a split-half reliability analysis must be considered later.

The correlation map is used to identify possible structures in the correlations and thus to quickly identify elements with interesting correlations.

Example

An example based on data collected from the Personality Tests website is permanently available on the XLSTAT Help Center. To download this data, go to:

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

References

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

Contingency tables (descriptive statistics)

Use this tool to compute a variety of descriptive statistics on a contingency table. A chi-square test is optionally performed. Additional tests on contingency tables are available in the "Tests on contingency tables" section.

In this section:

Description

A contingency table is an efficient way to summarize the relation (or correspondence) between two categorical variables V1 and V2. It has the following structure:

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

where $n(i,j) = n_{ij}$ is the frequency of observations that show both characteristic i for variable V_1, and characteristic j for variable V_2.

The Chi-square distance has been suggested to measure the distance between two categories. The Pearson chi-square statistic, which is the sum of the Chi- square distances, is used to test the independence between rows and columns. Is has asymptotically a Chi-square distribution with $(m_1 - 1)(m_2 - 1)$ degrees of freedom.

Inertia is a measure inspired from physics that is often used in Correspondence Analysis, a method that is used to analyse in depth contingency tables. The inertia of a set of points is the weighted mean of the squared distances to the center of gravity. In the specific case of a contingency table, the total inertia of the set of points (one point corresponds to one category) can be written as:

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

where n is the sum of the frequencies in the contingency table. We can see that the inertia is proportional to the Pearson chi-square statistic computed on the contingency table.

Bootsrap confidence intervals

XLSTAT allows you to obtain bootstrap confidence interval around the theoretical frequency of each pair of categories in a contingency table. It offers an alternative to the classical Chi-square by cell.

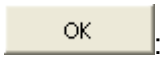
The method is as follow:

1. Build a dataset with two qualitative variables using the value of the contingency table.
2. Randomly draw with replacement N observations from the dataset for both variables independently.
3. Build a contingency table with the new dataset.
4. Repeat 2 and 3 as many times as specified by the user.
5. Compute mean, standard error, confidence interval and percentile confidence intervals for each pair of categories.

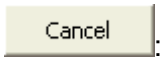
Pairs with observed value out of the confidence interval show a significant difference between the two categories (Amiri *et al*. 2011).

Dialog box

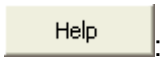
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Contingency table: Select the data that correspond to the contingency table. If row and column labels are included, make sure that the "Labels included" option is checked.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if the row and column labels are selected.

Options tab:

Chi-square test: Activate this option to display the statistics and the interpretation of the Chi-square test of independence between rows and columns.

Significance level (%): Enter the significance level for the test.

Bootstrap confidence interval: Activate this option to display the bootstrap confidence interval around the theoretical value for each pair of categories of the contingency table.

Number of samples: Enter the number of samples to be used to compute bootstrap confidence intervals.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Replace missing data by 0: Activate this option if you consider that missing data are equivalent to 0.

Replace missing data by their expected value: Activate this option if you want to replace the missing data by the expected value. The expectation is given by:

$$E(n_{ij}) = \frac{n_{i.}n_{.j}}{n}$$

$$E(n_{ij}) = \frac{n_{i.}n_{j.}}{n}$$

where $n_{i.}$ is the row sum, $n_{.j}$ is the column sum, and n is the grand total of the table before replacement of the missing data.

Outputs tab:

List of combines: Activate this option to display the table that lists all the possible combines between the two variables that are used to create a contingency table, and the corresponding frequencies.

Contingency table: Activate this option to display the contingency table.

Inertia by cell: Activate this option to display the inertia for each cell of the contingency table.

Chi-square by cell: Activate this option to display the contribution to the chi-square of each cell of the contingency table.

Significance by cell: Activate this option to display a table indicating, for each cell, if the actual value is equal (=), lower (<) or higher (>) than the theoretical value, and to run a test (Fisher's exact test of on a 2×2 table having the same total frequency as the complete table, and the same marginal sums for the cell of interest), in order to determine if the difference with the theoretical value is significant or not.

Observed frequencies: Activate this option to display the table of the observed frequencies. This table is almost identical to the contingency table, except that the marginal sums are also displayed.

Theoretical frequencies: Activate this option to display the table of the theoretical frequencies computed using the marginal sums of the contingency table.

Proportions or percentages / Row: Activate this option to display the table of proportions or percentages computed by dividing the values of the contingency table by the marginal sums of each row.

Proportions or percentages / Column: Activate this option to display the table of proportions or percentages computed by dividing the values of the contingency table by the marginal sums of each column.

Proportions or percentages / Total: Activate this option to display the table of proportions or percentages computed by dividing the values of the contingency table by the sum of all the cells of the contingency table.

Raw data: Activate this option to display the raw data table, meaning the observations/variables table, having N rows and 2 columns.

Charts tab:

3D view of the contingency table: Activate this option to display the 3D bar chart corresponding to the contingency table.

Contingency table: Activate this option to display the contingency table chart.

Proportions or percentages / Row: Activate this option to display the chart related to the *Proportions or percentages / Row* tab.

Proportions or percentages / Column: Activate this option to display the chart related to the *Proportions or percentages / Column* tab.

Chart options:

- **Chart type**
    - **Grouped**: Choose this option to display the graphs as bars grouped by modality.
    - **Stacked bars**: Choose this option to display the chart as stacked bars. These charts are used to compare the frequencies of sub-samples to those of a full sample.
- **Bar charts**
    - **Frequencies**: Choose this option to display the frequencies corresponding to each bar.
    - **Percentages**: Choose this option to display the % of population corresponding to each bar.

References

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.

Multiway crosstabs generator

Use this tool to create crosstabs (or cross tables) from as many categorical variables as needed.

In this section:

Description

Use this tool to create crosstabulations of categorical variables. You can use as many categorical variables as you want. A cross table is built by crossing qualitative variables which are found in lines, and others in columns. The cells of a crosstab can simply include the number of occurrences of a given cross in the dataset, the corresponding percentage, or a statistic calculated on a quantitative variable.

XLSTAT allows you to generate three types of crosstabs: * Either categorical variables are nested inside each other (for both rows and columns) * Either categorical variables are displayed one after the other (side by side) * Either you can generate all possible two-way crosstabs from the selected row and column variables (only two-way)

Example corresponding to the first format :

The original data consists of 3 qualitative variables (Age, Gender, Level of satisfaction). Age and gender are used as nested row variables, while level of satisfaction is used as a column variable.

|Age|Gender|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|

Example corresponding to the second format :

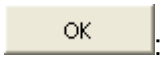
The original data consists, identical to the previous ones, of 3 qualitative variables (Age, Gender, Level of satisfaction). Age and gender are used as side-by-side row variables, while level of satisfaction is used as a column variable.

|||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| Gender|F|22|28|38|44|50| |M|15|26|24|39|58| ||

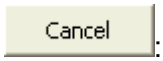
Statistics :

By default a crosstab makes it possible to calculate the number of cases corresponding to each cross as in the examples above, it is possible to calculate the % associated with the counts in the same way (by dividing the counts by the total number and multiplying by 100). Moreover, if quantitative variables are available, it is possible to calculate statistics for these variables, for each crossing. For example, in the case of a table that would cross on the one hand the age and the sex in rows, with the occupational category in column, one can display in the cells of the cross table the average income for the individuals corresponding to every possible crossing.

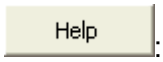
Dialog box 1



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

Data: Select the data you want to involve in the creation of the crosstab(s). If headers have been selected, check that the "Column labels" option is activated.

Weights: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column labels" option is activated.

Layout: Choose the display type for crosstabs. You can choose between the **Nested** display for which the variables are nested in the selection order (the order for row and column variables, is defined in the second dialog box), the **Side by side** display (the order is also defined in the second dialog box), or the **Two-way** display where the selected qualitative variables are all crossed pair by pair in two-way crosstabs.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

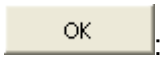
Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

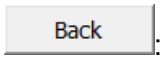
Column labels: Activate this option if the first row of the selected data (data, weights) contains a label.

Display the report header : Uncheck this option if you do not want to display the report header.

Dialog box 2



: Click this button to start the computations.



: Click this button to go back to the main the dialog box.

: Click this button to display the help.

Display totals: Activate this option to display marginal statistics, for rows and for columns.

Display sub-totals: Activate this option to display marginal statistics for each level of the penultimate row variable.

Sort categories: Activate this option so that within each qualitative variable, the categories are sorted in alphabetical order.

Variable-Category labels: Activate this option to use variable-category labels when displaying outputs. Variable-Category labels include the variable name as a prefix and the category name as a suffix.

Merge cells: Activate this option to merge the cells corresponding to the same variable (in lines or in columns).

Hide zeroes : Activate this option if you do not want to display null values in the crosstabs (typically crosses of values not encountered in the data).

One table per variable: Activate this option so that if statistics calculations are requested on several quantitative variables, the crosstabs are published for each variable separately.

One table per statistic: Activate this option so that if multiple statistics are requested, crosstabs are published for each statistic separately.

Row variables: Choose the variables to use for the crosstab rows. You can change the order by selecting a variable and then using the up or down arrows.

Column variables: choose the variables to use for the crosstab columns. You can change the order by selecting a variable and then using the up or down arrows.

Computations : In this block you can select: * **Quantitative variables** on which you want to compute statistics. If no variable is selected, you will only be able to compute counts and %. * **Statistics** that you want to compute for each crossing. The available statistics are: Counts, %, missing data, sum of weights, sum, mean, median, standard deviation. If other statistics would be useful to you, contact Addinsoft and they will be added. Note: Counts and sum of weights are identical if there are no missing data in a quantitative variable.

Example

An example showing how to create a crosstab is available on XLSTAT Help Center:

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

Ingelligent pivot tables

Use this module to turn an individuals/variables table into a dynamic pivot table optimized to let you understand and analyze the issue phenomenon corresponding to one of the variables describing the individuals.

In this section:

Description

"Ingelligent pivot tables" is a unique solution that allows you to quickly create intelligent pivot tables. This tool is based on classification trees using the CHAID algorithm in order to find the most relevant explanatory variables of a response variable.

A **pivot table** (or contingency table, or two-way table) is a synthetic representation of occurrences observed on an N-size population for crosses of all the different categories of two variables.

A **dynamic pivot table** allows to take more than two variables into account and to organize the table structure into a hierarchy. The table is said to be dynamic in the sense that software functionalities allow to navigate among the hierarchy and to create a focused view on particular classes of particular variables.

This tool allows you to create dynamic pivot tables whose structure is optimized with respect to a target variable. Numeric continuous or discrete explanatory variables are automatically sliced into classes that contribute to optimize the quality of the table.

The target variable can be a qualitative variable, or a quantitative variable.

This tool uses classification trees to discretize the quantitative variables and to identify the contributions of the variables (see the chapter on classification trees). The CHAID method is used because it suits well the dynamic pivot table representation.

When you run "Intelligent pivot tables" you will see successively two dialog boxes:

- The first dialog box lets you select the data and a few options.

- The second dialog box allows you to select the dimensions that you want to use in the pivot table (up to four variables may be selected). To help you select the variables the explanatory power of each variable are displayed. A specific score is used for that purpose (see bellow for a detailed description).

The function offers several options. However the default options should give the best results. For example, you can choose to use an external method for discretizing the quantitative explanatory variables. A sensitivity index is also available in order to better fit your needs in term of complexity of the tree generated.

Explanatory variables score index

In order to evaluate the contribution of the variables on the response variable, an index has been used. It will differ depending on the type of response variable.

In the case of a **quantitative response variable**, the score index for each variable (quantitative or qualitative), as defined by Breiman et al. (1984), is:

$$Score(var_i) = \sum_{j \in T} i^2 I(var_i \in Node_j)$$

With $i^2 = \frac{w_i \times w_j}{w_i + w_j}(\bar{y_i} - \bar{y_j})^2$, with i and j being the two nodes separating the studied node, T being the tree and $I(var_i \in Node_j) = \begin{cases} 1 \text{ if node j is associated to variable i} \\ 0 \text{ if not} \end{cases}$.

The weights w are computed with: $w_i = \frac{n_i}{N}(1 - \frac{n_i}{N})$, n_i being the number of observation associated to the leaf and N being the number of observations associated to the studied node.

In the case of a **qualitative response variable**, the score index for each variable (quantitative or qualitative), as defined by Breiman et al. (1984), is:

$$Score(var_i) = \sum_{j \in T} i^2 I(var_i \in Node_j)$$

With $i^2 = \frac{w_i \times w_j}{w_i + w_j} \sum_{k=1}^{nb\_mod} (p_{ik} - p_{jk})^2$ with i and j being the two nodes separating the studied node, T being the tree and $I(var_i \in Node_j) = \begin{cases} 1 \text{ if node j is associated to variable i} \\ 0 \text{ if not} \end{cases}$.

The weights w are computed with: $w_i = \frac{n_i}{N}(1 - \frac{n_i}{N})$, n_i being the number of observation associated to the leaf and N being the number of observations associated to the studied node. The probabilities are the probabilities of having modality k of the response variable in each leaf.

Sensitivity index associated to the tree

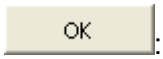
Building a classification tree requires to set a number of parameters (maximum depth, leaf size, the thresholds of grouping and separation ...). To simplify the use of "Intelligent pivot tables", a sensitivity index was developed. It takes values between 0 and 1.

When this index is close to 0, then the building of the tree is not sensitive to small differences. The number of intervals in the discretization of the quantitative variables will be lower and the size of the tree will be small. It is therefore the strongest contributions that will be revealed in the pivot table.

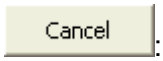
When this index is close to 1, then the building of the tree is very sensitive to small differences. The number of intervals in the discretization of the quantitative variables will be larger and the size of the tree will be large. All contributions will be revealed in the pivot table (but sometime too many).

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Response variable: Select the response variable you want to model. If a column header has been selected, check that the "Variable labels" option has been activated.

Choose the type of response variable you have selected:

- **Quantitative**: If you select this option, you must select a quantitative variable.

- **Qualitative**: If you select this option, you must select a qualitative variable. You must then select a target category which will be used for the outputs of the pivot table. A new box with the list of the categories of the response variable will appear on the right.

X / Explanatory variables

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If a variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If a variable header has been selected, check that the "Variable labels" option has been activated.

Weights: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Sample labels" option is activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (response and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Options tab:

Sensitivity: Enter the value of the sensitivity parameter. When it is close to 1, the classification tree is large. When it is close to 0, the classification tree is small. For a detailed description, please refer to the description part of this chapter. The default value is 0.5.

Discretization – quantitative variables: this option is enabled only if quantitative explanatory variables have been selected.

- **Automatic**: Activate this option to use the automatic discretization within the tree algorithm (this is the default option).
- **Equal width**: Activate this option to discretize the quantitative variable using equal width intervals.
- **Equal frequency**: Activate this option to discretize the quantitative variable using equal frequency intervals.
- **User defined**: Activate this option to discretize the quantitative variable using user defined interval. Select a table with one row for each bound of the intervals and one column for each variable.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Discretization: Activate this option to display the discretized explanatory variables.

Contributions: Activate this option to display the contributions table and the corresponding bar chart.

Pivot table: Activate this option to display the dynamic pivot table.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables. For qualitative variables, including the response variable, the categories with their respective frequencies and percentages are displayed.

The next table presents the **discretized explanatory variables**.

The next table presents the **variables contributions** (raw, % relative and cumulated contribution). This table allows you to quickly see which variables have the greater impact on the target variable. A bar chart of the contributions is also displayed. This histogram is an Excel chart that you can modify to suit your needs.

The most important result is the **dynamic pivot table**. Each cell corresponds to a unique combination of the values of the explanatory variables. It is described by the following 4 values, that can be displayed or not according to the user preferences:

- **Target average**: Percentage of the cases where the target category of the response variable is present in the case of a qualitative variable; average of the target variable calculated on the sub-population corresponding to the combination in the case of continuous variable;
- **Target size**: Count of the occurrences of the target category for the response variable in the case of qualitative variable;
- **Population size %**: Percentage of the overall population corresponding to the combination;
- **Population size**: Population size corresponding to the combination.

Example

An example based on data collected for a population census in the United States is permanently available on the XLSTAT Help Center. To download this data, go to:

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

References

Breiman, L. , Friedman, J.H., Olshen, R. A. and Stone, C.J. (1984) . Classification and regression tree, Chapman & Hall.

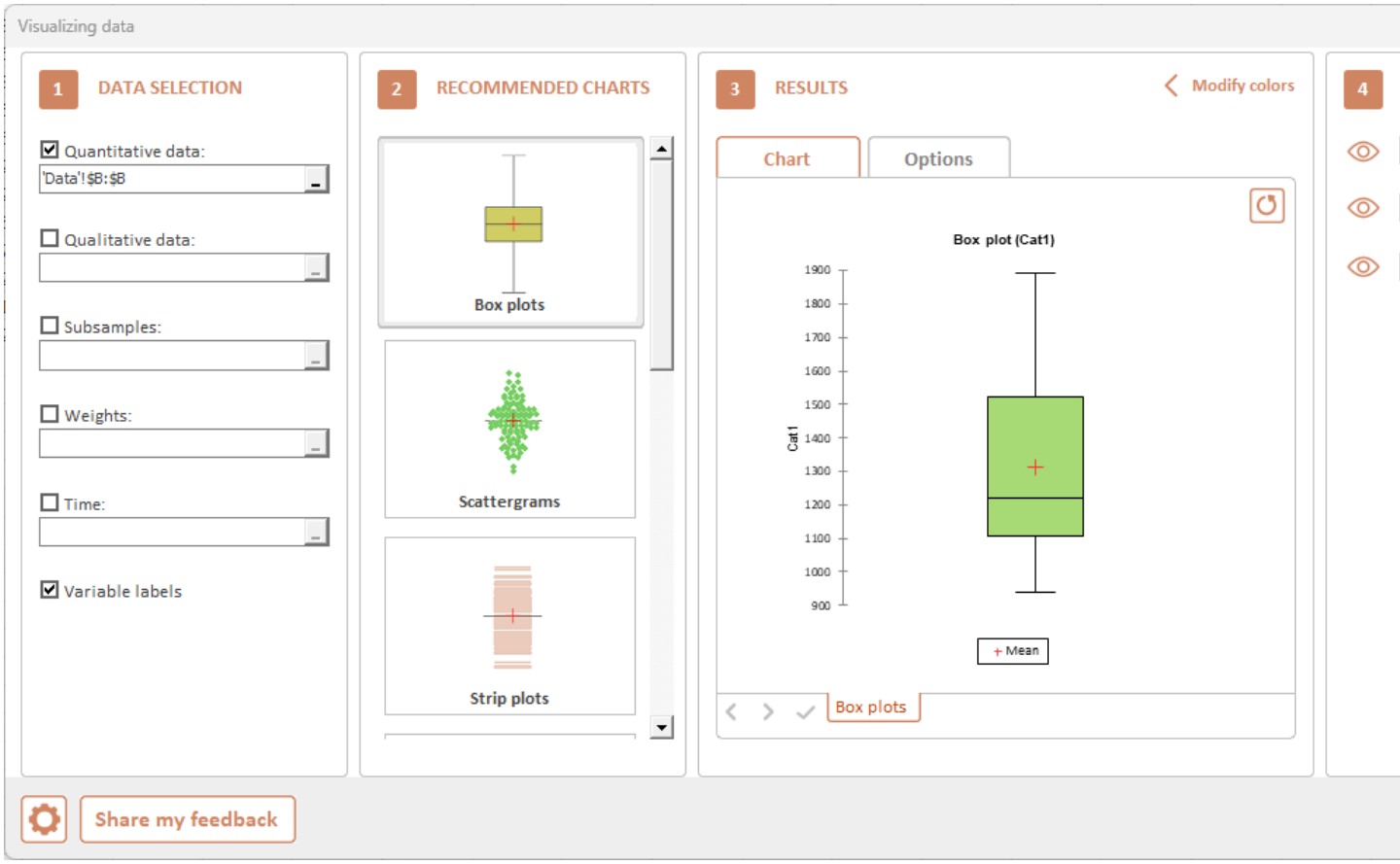
Visualizing data

DataViz

Use DataViz to find the ideal graphic for your needs in just a few clicks and customize it with ease.

In this section:

Generate a customizable graph in just a few clicks



DataViz is the perfect tool for selecting a graph that matches your data without having to navigate between several features.

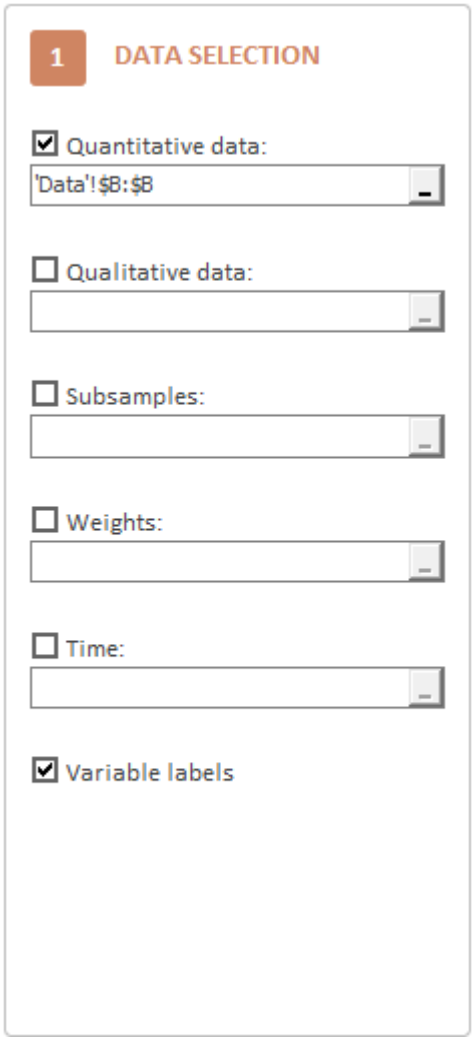
With DataViz, you can generate graphs in just 4 steps.

The chosen chart can be customized and exported to Excel for integration into your analyses and reports.

Explore the rich functionality of DataViz

1. Select data to view

Select data to view



The Dataviz tool enables you to select any type of data to generate an appropriate visualization:

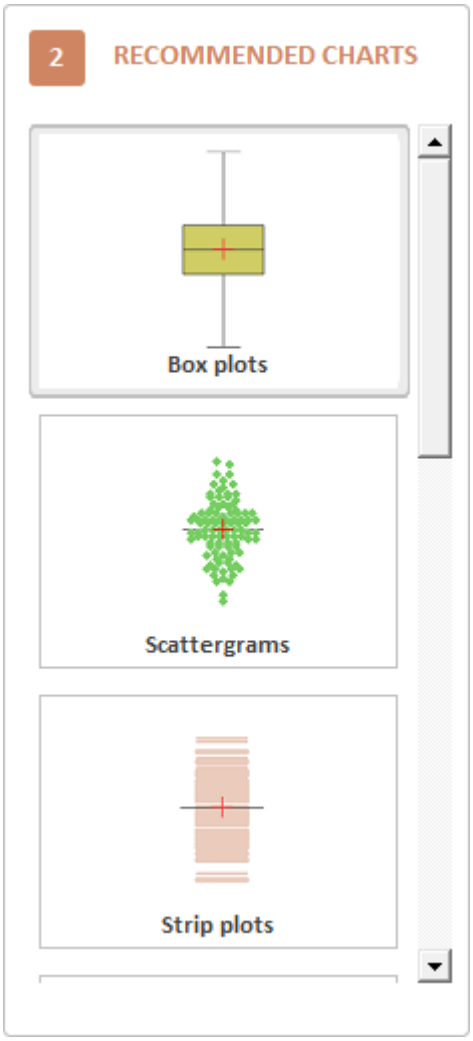
- quantitative data,
- qualitative data,
- subsamples,
- weights
- time.

An automatic detection function warns you of any incompatibility between the selected data and the chosen data type.

Indicate whether the data supplied includes a header.

2. Choose the chart that best illustrates the data

List of recommended charts



A list of recommended graphics is displayed, showing all the visualization possibilities available.

The thumbnail of each type of graph makes it easy to select the one that will intuitively highlight the desired result.

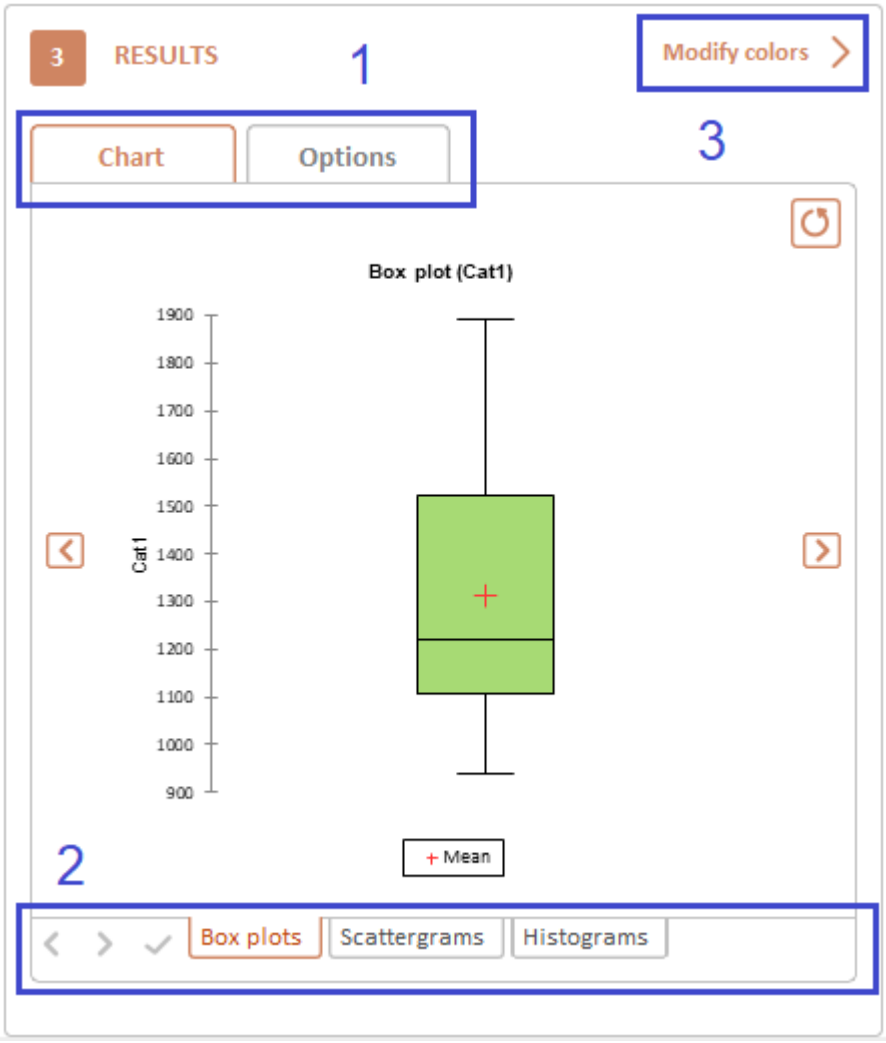
Benefit from a very quick update of the chart list, after a change in data selection.

Click on  once you've chosen the type of graph you want to generate.

Note: As DataViz is still in beta version, it may not offer charts for all combinations of data types. However, new graphs will be progressively integrated in future versions.

3. Chart settings

Chart settings



Once the graph has been generated, DataViz offers a range of dynamic actions.

Navigate between the different graphs generated using the chevrons.

This feature is particularly useful, for example, when creating a category chart.

Simply click on to return to the initial chart, before modifying any options or colors.

If a time variable is present (**Time** field), interact with the graph using the button panel.

These buttons allow you, in order, to:

- Go back in time to grasp the evolution of the graph.
- Pause playback.
- Read the graph's evolution over time.
- Speed up or slow down playback.

Simply click on a chart title or axis to modify it.

View a chart or redefine chart-specific parameters by navigating between the **Chart** and **Options** tabs (Zone 1). Experiment and regenerate parameters as you wish, for results that reflect your own image.

Explore the history of generated graphics with ease using the < > (*Zone 2*). Drag tabs left or right when they're no longer visible, or return instantly to the current tab by clicking on ✓ (*Zone 2*).

In each history tab, you'll find a graph generated for a "selected data - recommended graph" set. The active tab will be colored orange.

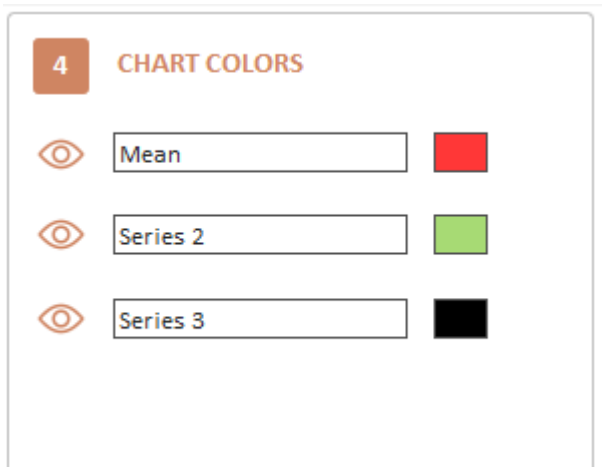
To modify a tab, simply click on its name. To remove it, right-click on its name.

Finally, customize the chart colors by clicking on Modify colors > (*Zone 3*).

Note: Modifying options will not create a new tab.

4. Customize chart colors

Click on Modify colors > to view and customize the series in your graphics.



Customize colors

The Dataviz tool lets you select any type of data to generate an appropriate visualization: each series is represented by a line, offering three interactive options:

- Get an immediate overview of the series in the generated chart with a single click on 👁.
- Easily rename a series via the respective text box, without any impact on the chart. This makes it much easier to identify each series, especially when a large number are used.
- Express your creativity by choosing one or more new colors, by clicking on the color square. Thanks to the color selector, you can choose from a wide range of shades.

5. Export the customized chart for optimum use in Excel

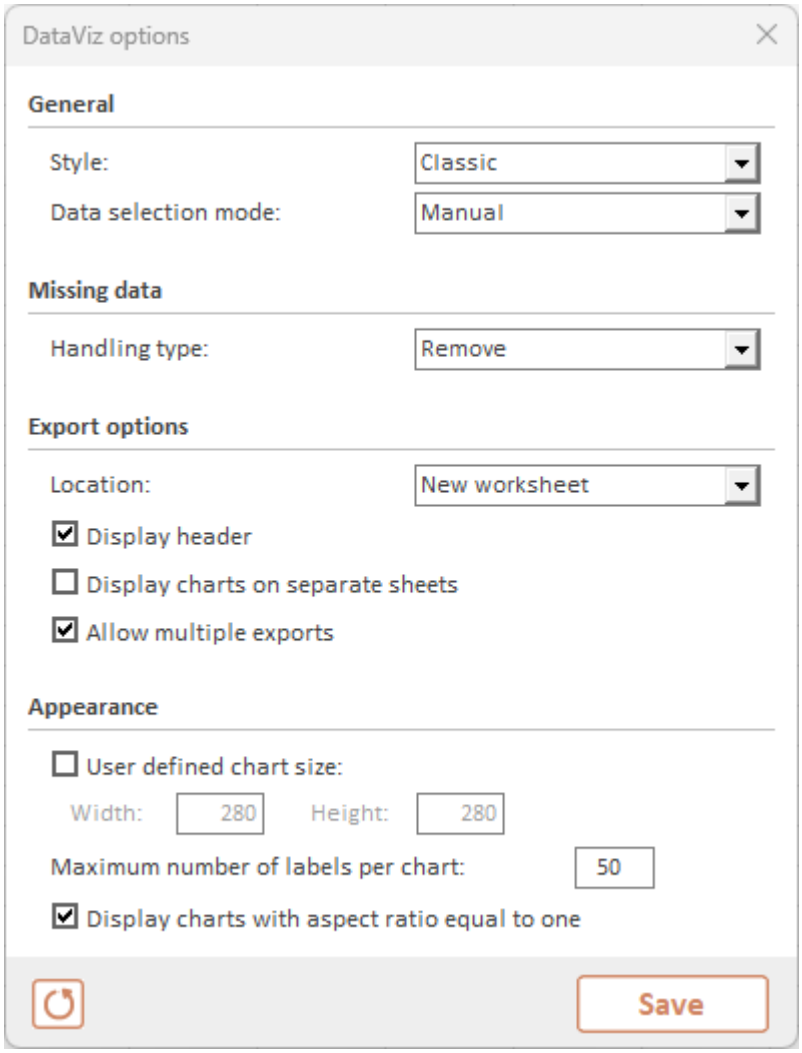
Once you've created your chart, all that's left to do is export it.

Simply click on the 

icon in the bottom right-hand corner of the window, and you'll be able to export the chart(s) in the current tab.

For further details, please refer to the general options dialog box.

General options dialog box



- **General**:
  - **Style**: This option lets you modify the colors of the generated chart. Choose the style you prefer from "Classic", which corresponds to XLSTAT's historical format, "Modern", which corresponds to another color scheme, and "Scientific", which uses only black, white and grayscale. This option does not apply to all charts.
  - **Data selection mode**:
    - **Manual**: choose this option if you wish to select data directly from an Excel sheet.
    - **Variables**: choose this option to display a dialog box with the list of data automatically retrieved from the active sheet. All you have to do is tick the data to be used as input.

- **Missing data**:
    - **Handling type**: choose how the missing data should be processed.
        - **Refuse**: Choose this option to not generate a chart when data is missing.
        - **Remove**: Choose this option to delete missing data before generating a chart.
        - **Replace**: Choose this option to replace missing data with its average before generating a chart.

  Note: Some charts do not yet support missing data. This setting therefore does not apply to them.

- **Export options**
    - **Location**:
        - **New worksheet**: Choose this option to export charts to a new worksheet in the active workbook.
        - **New workbook**: Choose this option to export charts to a new workbook.
        - **Custom location**: Choose this option to select a cell for the export location.
    - **Display header**: Activate this option to display a header in export sheets. The header includes buttons for relaunching the exported chart in DataViz and for exporting charts in the sheet to Word or Powerpoint.
    - **Display charts on separate sheets**: Activate this option to have charts exported on separate chart sheets.

  Note: When a chart is displayed on a standard Excel sheet, you can convert it to a separate chart sheet by following these steps: - Select the chart. - Right-click on "Move Chart". - Choose "New worksheet".

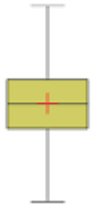
- **Allow multiple exports**: activate this option to export multiple charts before closing the DataViz dialog box.

- **Appearance**:
    - **User-defined chart size**: Activate this option if you want XLSTAT to display charts whose size is exactly defined by the values below.
    - **Width**: Enter the value in points for the width of the charts.
    - **Height**: Enter the height of the charts in points.
    - **Maximum number of labels per chart**: Enter the maximum number of labels to be displayed on a chart.
    - **Display charts with aspect ratio equal to one**: Activate this option to display factorial analysis charts as orthonormal. This automatically ensures identical abscissa and ordinate scales, and avoids misinterpretations due to artificial dilation effects.

List of recommended charts

Recommended charts for quantitative data

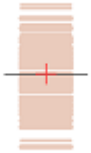
Box plot



Ideal for **quantitative data**, you can include **subsamples** and add **weights** to the data.

Also known as a "moustache box", it's a rectangle displaying the minimum, 1st quartile, median, mean, 3rd quartile, as well as the two limits beyond which values can be considered abnormal. The mean is displayed as a red +, and the median as a black line.

Strip plot



Ideal for **quantitative data**, you can include **subsamples** and add **weights** to the data.

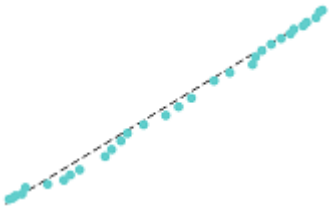
Represents sample data in strip form. Over a given interval, the tighter or thicker the bands, the more data there is.

Scattergram

Ideal for **quantitative data**, you can include **subsamples** and add **weights** to the data.

Gives an idea of the distribution and possible plurality of modes of a sample. All points are shown, along with the mean and median.

Normal P-P plots



Ideal for **quantitative data**, you can include **subsamples** and add **weights** to the data.

Used to compare the empirical distribution function of a sample with that of a sample distributed according to a normal distribution with the same mean and variance. If the sample follows a normal distribution, the points must coincide with the first bisector of the plane.

Normal Q-Q plots

Ideal for **quantitative data**, you can include **subsamples** and add **weights** to the data.

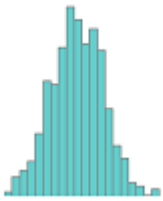
Compares the quantiles of the sample with those of a sample distributed according to a normal distribution with the same mean and variance. If the sample follows a normal distribution, the points must coincide with the first bisector of the plane.

Means chart

Ideal for **quantitative data**, you can include **subsamples** and add **weights** to the data.

Shows averages for each variable, in the form of bar charts. It is also possible to display the **error bars** in three different forms.

Histogram



Ideal for **quantitative data**, you can include **subsamples** and add **weights** to the data.

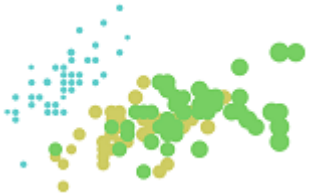
Gives you a quick idea of the distribution of a sample of continuous or discrete quantitative data.

Correlation tests

Ideal for **quantitative data**, it requires at least **2 qualitative data**. You can include **sub-samples** and add **weights** to the data.

Represents a correlation matrix whose values have been colored according to a customizable scale. This map lets you see at a glance which variables are highly correlated.

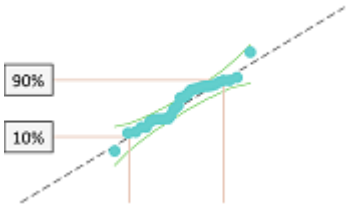
Scatter plot



Ideal for **quantitative data**, you can include **subsamples** and add **weights** to the data.

Represents data in 2 or 3 dimensions.

Probability plot



Ideal for **quantitative data**, you can include **subsamples** and add **weights** to the data.

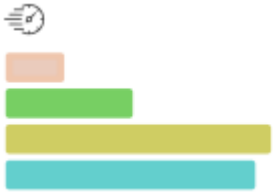
Helps to visually check whether a sample comes from a population with a given distribution.

Tornado diagram

Ideal for a variable containing **quantitative data** accompanied by **subsamples** or for **2 quantitative data**.

Similar to a bar chart, it compares the relative importance of two variables.

Bar chart race



Ideal for modeling **quantitative data** accompanied by **time data**, you can include **subsamples**.

The bar chart race visualizes the evolution of a variable over time on a single dynamic chart.

Motion Chart

Ideal for modeling **2 quantitative data** accompanied by **time data**. You can include **subsamples**.

Display the evolution of several variables (up to 3), measured on several individuals, over time on a single dynamic chart.

Contour plot

Suitable for **quantitative data**, it requires **3 quantitative columns** to be used.

Used to study the relationship between a response variable and two prediction variables.

Recommended charts for qualitative data

Pie chart

Ideal for **qualitative data**, you can include **subsamples** and add **weights** to the data.

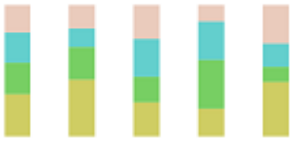
Represents, in the form of pie charts, the numbers or frequencies of the different modalities of qualitative variables.

Bar chart

Ideal for **qualitative data**, you can include **subsamples** and add **weights** to the data.

Represents, in the form of bar charts, the numbers or frequencies of the different modalities of qualitative variables.

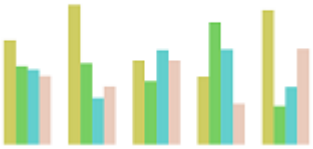
Stacked bar



Suitable for **qualitative data** subdivided into **subsamples**. You can include **weights** to the data.

Compares the numbers or frequencies of subsamples with those of a full sample.

Clustered bar



Suitable for **qualitative data** subdivided into **subsamples**. You can include **weights** in the data.

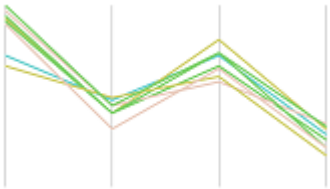
Compares the numbers or frequencies of subsamples with those of a full sample.

2D plots for crosstabs

Ideal for **qualitative data**, it requires at least **2 qualitative data** to be processed.

Generates a 2D chart showing the relative importance of the various combinations you can obtain when creating a contingency table or, more generally, a pivot table.

Recommended charts for mixed data

Radar charts	parallel coordinates plots	Semantic differential chart
		
Suitable for **quantitative data**, it requires **qualitative data** to be used. Evaluate different choices according to several variables.	Suitable for **quantitative** and **qualitative** data, you can include **subsamples** and add **weights** to the data. Display multi-dimensional data on a single two-dimensional chart.	Suitable for **quantitative** and **qualitative** data. View scores assigned by subjects to objects for different criteria.

Example

A tutorial is available on the XLSTAT Help Center:

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

Probability plots

Use this tool to create probability plots that allow to visually control if a sample may come from a population that follows a given distribution.

In this section:

Description

Probability plots is an old method (Hazen, 1914), that has been extensively used, especially through the use of printed probability paper. It is useful to visually control whether a sample follows a given distribution.

Any XLSTAT distribution can be used (see the Histogram tool for the full list). The fitting of the distribution can done before creating the plot. The distribution that fits best can be automatically chosen by XLSTAT, or you can select a specific distribution and choose to enter the parameters or let XLSTAT estimate them.

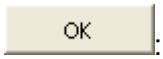
Let $\{x_1, x_2, ..., x_n\}$ be the order statistics of a sample of size n that is supposed to follow a distribution $F(x)$. To construct a probability plot, x_i is plotted against $F^{-1}(p_i)$, where p_i is the estimate of $F(x_i)$, namely the plotting position. Several approaches have been proposed to compute p_i. XLSTAT includes the following options:

- 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): $p_1 = 1 - 0.5^n/p_n = 0.5^n/p_i = (i - 0.3175)/(n + 0.365)$ $(1 < i < n)$.

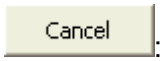
XLSTAT proposes as well the possibility to compute the estimated order statistics for a given distribution using Monte Carlo simulations. Monte Carlo simulations are also used by XLSTAT to compute confidence intervals. Asymptotic confidence intervals are also available (Chambers, 1983).

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data: Select the data for which you want to compute and display a probability plot. If a sample header has been selected, check that the "Sample labels" option has been activated.

Weights: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Sample labels" option is activated.

Subsamples: Check this option to select a column showing the names or indexes of the sub-samples for each of the observations.

- **Colors only**: Activate this option if the subsamples information should only be used to color data on the probability plot. Otherwise it will be used to compute a separate probability for each subsample.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Sample labels: Activate this option if the first row of the data selections (data, weights, subsamples) includes a header.

Options tab:

Distribution: Choose the probability distribution to be used for the fit and/or goodness of fit tests. See the description section of the distribution fitting function, for more information on the distributions offered. The **automatic** option allows to let XLSTAT identify the best fitting distribution (determined using a Kolmogorov-Smirnov test).

Parameters: You can choose to **enter** the parameters for the distribution, or **estimate** them. If you choose to enter the parameters, you must enter their values.

Estimationmethod: Choose the method of estimating the parameters of the chosen distribution.

- **Moments**: Activate this option to use the moments method.
- **Maximum likelihood**: Activate this option to use the maximum likelihood method. You can then change the **convergence** limit value which when reached means the algorithm is considered to have converged. Default value: 0.00001.

Method: Select the method you want to use to estimate the plotting positions (see the description section for more information). If the method is Monte Carlo, you can then specify the number of simulations and the maximum time to spend on the simulations.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations:

- **For the corresponding sample**: Activate this option to ignore an observation which has a missing value only for samples which have a missing value.
- **For all samples**: Activate this option to ignore an observation which has a missing value for all selected samples.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Charts tab:

Log scale: Activate this option to use a log scale on the probability plot.

Abscissa=Observed: Activate this option so that the observed values correspond to the abscissa axis. Otherwise they correspond to the ordinates.

Display %: Activate this option to use the cumulative % as a scale for the axis used for the plotting positions.

Confidence intervals: Activate this option to display the confidence intervals. The value you enter (between 1 and 99), in **percentage**, is used to determine the confidence intervals for the estimated values. The default value is 95.

Percentiles: Activate this option and select up to four percentiles that you want to display on the chart.

Quantiles: Activate this option and select up to four quantiles that you want to display on the chart.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables. For qualitative variables, including the response variable, the categories with their respective frequencies and percentages are displayed.

If a distribution has been fitted to the data, the results of the fit and the estimated parameters of the distribution are displayed.

The probability is then displayed, with, if the corresponding option has been selected in the dialog box, the confidence intervals.

Example

An example showing how to create probability plots with XLSTAT is available at:

www.xlstat.com/demo-probaplots

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.

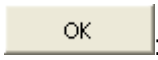
Scatter plots

Use this tool to create 2- or 3-dimensional plots (the 3rd dimension being represented by the size of the point), or indeed 4-dimensional plots (a qualitative variable can be selected). This tool is also used to create matrices of plots to enable a study of a series of 2-dimensional plots to be made at the same time.

Note: XLSTAT-3DPlot can create plots with much more impact thanks to its large number of options with the possibility of representing data on a third axis.

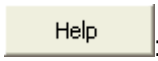
In this section:

Dialog box



: Click this button to start the calculations.

: Click this button to close the dialog box without doing any calculations.



: Click this button to display help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

X: In this field select the data to be used as coordinates along the X-axis.

Y: In this field select the data to be used as coordinates along the Y-axis.

Z: Check this option to select the values which will determine the size of the points on the charts.

- **Use bubbles**: Check this option to use charts with MS Excel bubbles.

Groups: Check this option to select the values which correspond to the identifier of the group to which each observation belongs. On the chart, the color of the point depends on the group.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selected data (X, Y, Z, Groups, Weights and observation labels) contains a label.

Observation labels: Check this option if you want to use the available line labels. If you do not check this option, labels will be created automatically (Obs1, Obs2, etc.). If a column header has been selected, check that the "Variable labels" option has been activated.

Weights: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Matrix of plots: Check this option to display all possible combinations of variables in pairs in the form of a two-entry table with Y-variables in rows and Y-variables in columns.

- **Histograms**: Activate this option so that if the X and Y variables are identical, XLSTAT displays a histogram instead of a X/X plot.
- **Q-Q plots**: Activate this option so that if the X and Y variables are identical, XLSTAT displays a Q-Q plot instead of a X/X plot.

Frequencies: Check this option to display the frequencies for each point on the charts.

- **Only if >1**: Check this option if you only want frequencies strictly greater than zero to be displayed.

Confidence ellipses: Activate this option to display confidence ellipses and choose the size of the corresponding confidence interval. The calculation of the ellipse can be done using either the Fisher or the Chi-square distribution.

Legend: Check this option if you want the chart legend to be displayed.

Trend lines: Activate this option to display trend lines on the chart. Several types of trend lines are available in XLSTAT: * Regression lines * Polynomial * Exponential * Logarithmic * Cubic spline

Tab **Colors** :

Number of groups : Activate this option to choose a color for each group. Then enter the number of modalities of the variable Groups selected in the General tab.

Group : Click on each group field to choose the respective color.

Example

A tutorial on using Scatter plots is available on XLSTAT Help Center:

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

References

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.

Motion charts

Use this tool to explore the evolution of several variables over time.

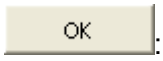
In this section:

Description

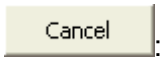
Usually, when one wants to visualize the evolution of a variable over time, the value of the variable is plotted on the Y axis as a function of time on the X axis. With this type of representation, you are limited to the evolution of only one variable at a time. If more than one variable on different observations are to be explored, several charts are needed. The XLSTAT Motion Chart solves this issue by allowing you to explore the evolution of several variables (up to 3) measured on different observations over time on a single dynamic chart.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the

arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

X: In this field select the data to be used as coordinates along the X-axis.

Y: In this field select the data to be used as coordinates along the Y-axis.

Time: In this field select the date or time data, or any numerical index corresponding to the observations of the time series.

Groups: In this field select the data that indicate the group to which each observation belongs.

Size: Check this option to select the values which will determine the size of the points on the charts.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selected data (X, Y, Time, Groups and Size) contains a label.

Interpolate missing points: Check this option to estimate coordinates of missing points as the mean of previous and next coordinates.

Smoothing: Check this option to display intermediate positions of points between two consecutive dates in order to produce a smoother dynamic evolution between each time steps.

Legend: Check this option to display a legend with the minimum and maximum value of Size according the size of points.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove the observations with missing data.

Replace by the average of the previous and next values: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

Buttons on Motion Chart

: The scrollbar at the bottom of the chart allows to change values of the time variable and thus display positions of points at different times.

: Click this button to automatically decrease values of the time variable.

: Click this button to stop the moving of points on the chart.

: Click this button to automatically increase values of the time variable.

: Click this button to set the speed of movement (from 1 very slow to 10 very fast) of the points over time.

Example

A tutorial on generating motion chart is available on the XLSTAT Help Center at the following address:

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

Bar chart race

Use this tool to visualize the evolution of a quantitative variable over time for several groups of observations.

In this section:

Description

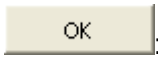
To visualize values of a variable for different groups (categories), we often use bar charts. The height of each bar represents the value of the measured variable for a given group. However, if the variable is measured at different times, the classical bar chart cannot be used to show the evolution over time. This is where the bar chart race comes in handy. It allows to visualize the evolution of a variable over time on a single dynamic chart.

Two options are available. You can choose to represent the raw data at each time interval, or display the cumulated data. Display of cumulated data can be very useful for counting data for example.

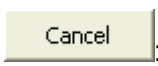
If for a time t there is no data for a given group, then the displayed value will be the value of the group at the previous time.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Data: Select the quantitative variable you want to visualize on the bar chart.

Time: Select the dates or time data, or any numerical variable corresponding to the time series.

Groups: Select the group identifiers to which each observation belongs.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selected data (Data, Time, Groups) contains a label.

Options tab:

Position of the bars:

- **Fixed position**:
    - **Smallest at bottom**: The group with the lowest value at the last time will be displayed at the botton of the chart.
    - **Largest at bottom** : The group with the highest value at the last time will be displayed at the bottom of the chart.
    - **Sort alphabetically** : Groups are displayed in alphabetical order.
- **Colors**:

- **Distinct colors**: Each group is displayed with a different color.
- **Color scale**: A graduated color scale is used so that the group with the highest value at the last time is displayed in dark red.

- **Data**:
  - **Cumulate**: At each time interval t the displayed value is the value present in your data at time t plus the value at time $t-1$.
  - **Raw data**: At each time interval the displayed value corresponds to the value present in your data.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Use previous value: Activate this option to replace a missing data at time t with the present value at time $t-1$.

Buttons on bar chart race

: The scrollbar at the bottom of the chart allows to change the value of the time variable and thus display bars at a specific date.

: Click this button to move backward in time.

: Click this button to stop the bar chart race animation.

: Click this button to move forward in time.

: Click this button to set the speed of the animation (1 very slow - 10 very fast).

Example

A tutorial on generating bar chart race is available at the XLSTAT Help Center at the following address:

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

Parallel coordinates plots

Use this tool to visualize multidimensional data (described by p_1 quantitative and p_2 qualitative variables) on a single two dimensional chart.

In this section:

Description

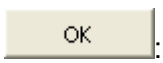
This visualization method is useful for data analysis when you need to discover or validate groups. For example, this method could be used after Agglomerative Hierarchical Clustering (AHC).

If you consider n observations described by p_1 quantitative and p_2 qualitative variables, the chart consists of $p = p_1 + p_2$ vertical axes each representing a variable, and n lines corresponding to each observation. A line crosses an axis to the value the observation takes on the corresponding variable.

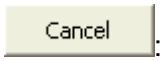
Because of the Excel restrictions (maximum of 255 data series) you cannot represent more than 255 observations. If you select more observations, only the first 255 will be displayed. Also when many observations are present, the graph can quickly become unreadable. This is why XLSTAT allows you to represent descriptive statistics summarizing information instead of representing all observations.

Dialog box

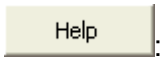
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Quantitative Data: Activate this option to select quantitative data describing the observations. If a column header has been selected, check that the "Variable labels" option is activated.

Qualitative Data: Activate this option to select qualitative data describing the observations. If a column header has been selected, check that the "Variable labels" option is activated.

Weights: Weights can be only applied when you have chosen the Display statistics lines option (see below) in the Options tab. If you do not activate the Weights option, the weights will all be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Groups: Check this option to select the values which correspond to the identifier of the group to which each observation belongs. If a column header has been selected, check that the "Variable labels" option is activated. On the chart, the color of the point depends on the group.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selected data (quantitative data, qualitative data, weights, groups and observation labels) contains a label.

Observation labels: Check this option if you want to use the available line labels. If you do not check this option, line labels will be created automatically (Obs1, Obs2, etc.). If a column header has been selected, check that the "Variable labels" option has been activated.

Options tab:

Display type:

- **One line per observation**: Activate this option to display as many lines as possible (the maximum is 250 due to the limitations of Excel).
- **Display statistics lines**: Check this option to display a line for each of the following statistics:
- **Minimum and maximum**
- **Median**
- **First quantile (%)**: Enter the value of the first quantile (2.5% by default).
- **Second quantile (%)**: Enter the value of the second quantile (97.5% by default).
- **Mode** (for qualitative variables)

Quantitative data:

- **Raw data**: Activate this option to use the raw data of the quantitative variables. The scale on the Y-axis will be between the minimum and maximum of all variables.
- **Rescale**: Activate this option so that all variables are represented on the same scale between 0 (minimum) and 1 (maximum).
- **Different Y axes**: Activate this option so that each Y-axis has a scale adapted to the associated variable.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Ignore missing data: Activate this option to ignore missing data. If missing data are present they will not be displayed on the chart.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Example

A tutorial on generating a parallel coordinates plot is available on the XLSTAT Help Center at the following address:

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

References

Inselberg A. (1985). The plane with parallel coordinates. *The Visual Computer*, **1**, pp. 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, pp 664-675.

Ternary diagrams

Use this tool to create ternary diagrams to represent within a triangle a set of points that have their coordinates in a three-dimensional space, with the constraint that the sum of the coordinates is constant.

In this section:

Description

This visualization method is particularly useful in domains where one works with three elements with varying proportions, for example in chemistry or petrology.

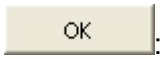
This tool lets you quickly create a ternary diagram representing points and the projection lines connecting each point to each axis.

There are two approaches for ternary graphs:

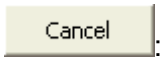
- Either segments corresponding to the orthogonal projection of the points on the axes give the information on the relative proportions of the three elements.
- Or the projection parallel to the axis A onto the axis B corresponds to the coordinate of the point along the axis B, where B is after A when turning counterclockwise.

XLSTAT currently allows only the second approach.

Dialog box



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the

arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

X: Check this option to select the data corresponding to the first element.

Y: Check this option to select the data corresponding to the second element.

Z: Activate this option to select the quantitative data that correspond to the third element.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selected data contains a label.

Options tab:

Constant: Check this option to display as many lines as possible (the maximum is 250 due to the limitations of Excel).

Charts tab:

X1/X2 | Min / Max: You can modify the min and the Max for each dimension. However you must take into account that the (Max-Min) for each dimension is the same. The min and Max for the third dimension is automatically computed from the min and Max for the first two dimensions.

Number of segments: Enter the number of segments into which you want to divide each axis of the ternary chart.

Projection lines: Activate this option to display dotted red lines between the points and their coordinate on each axis.

Lines between axes: Activate this option to display the lines between the axes.

Link to input data: Activate this option to link the chart to the input data. If you check this option, a change in the input data is immediately reflected on the ternary diagramm.

Example

A tutorial on generating ternary plots is available on the XLSTAT Help Center at the following address:

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

2D plots for crosstabs

Use this tool to create a 2-dimensional plot based on a contingency table or, more generally, on a crosstab.

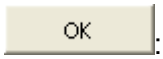
In this section:

Description

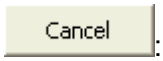
This visualization tool allows to quickly generate a 2D plot showing the relative importance of values contained in a two-way contingency table (a contingency table displays the counts for the different combinations of the categories of two qualitative variables - for example, the counts of respondents grouped into age groups that replied to a yes/no question) or more generally in a crosstab (for example, the total of sales by product in different countries).

This tool can work directly on raw data (weighted or not) or on a cross tab.

Dialog box



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data format: Select the data format.

- **Crosstab**: Activate this option if your data correspond to a contingency table or a crosstab.
- **Qualitative variables**: Activate this option if your data are available as two qualitative variables to be used to create a contingency table.

Contingency table: If the data format is "contingency table", select the data that correspond to the contingency table. If row and column labels are included, make sure that the "Labels included" option is checked.

Qualitative variable(1): If the data format is "qualitative variables", select the data that correspond to the qualitative variable that will be used to construct the rows of the contingency table, and that will be used for the ordinates axis of the plot.

Qualitative variable(2): If the data format is "qualitative variables", select the data that correspond to the qualitative variable that will be used to construct the columns of the contingency table, and that will be used for the abscissa axis of the plot.

Z: If the data format is "qualitative variables", check this option to select the values which will weigh the observations and modify the size of the points on the plot. If you want to display several dimensions on the plot, you can select several columns. If you want that the same scale is used for each dimension on the plot, activate the **same scale** option.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Labels included: Activate this option if the row and column labels of the contingency table are selected.

Variable labels: Activate this option if the first row of the data selections includes a header.

Display title: Check this option, to display a title on the plot.

Options tab:

Shape: Select the shape you want to use.

- **Circle**
- **Square**
- **Bubbles**

Show values: Activate this option to display the values on the plot.

Size: Choose which dimension is related the values (Z): area, width or width². With area, the area of the points is proportional to the values, so this is diminishing the difference between small and large values. It is the opposite with width².

Scale(%): Choose how the points should be rescaled. The default is 100 and corresponds to no rescaling.

Horizontal axis at bottom: Check this option to display the horizontal axis at the bottom of the chart.

Gridlines: Check this option to display gridlines on the chart.

Example

A tutorial on generating a 2D plot for a contingency table is available at:

http://www.xlstat.com/demo-2dcont.htm

Error bars

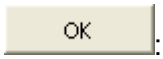
Use this tool to easily create Excel charts with error bars that can be different for each point.

In this section:

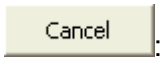
Description

This tool has is there to get around a failure of Excel: if it is possible to add error bars on different types of graphs, this operation is tedious if the bounds are not the same for all points. With this tool you can create a chart with error bars at once.

Dialog box



: Click this button to create the charts.



: Click this button to close the dialog box without creating the charts.

: Click this button to display help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables (column mode). If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations (row mode).

General tab:

X: Select in this field the data to be used as coordinates for the x-axis. If you select several columns (column mode) or several rows (row mode), you must then select the same number of columns (or rows) for Y, the lower bounds, and the upper bounds. However, if you select a single column (or row), you can then select one or more columns (or rows) for Y, the lower bounds and upper bounds

Y: Select in this field that data to be used as coordinates on the y-axis. See above the constraints that apply to the number of columns.

Lower bound: Activate this option if you want to add lower bounds on the graph. Then select in this field the data to be used as lower bounds. The number of columns (column mode) or rows (row mode) to be selected must be equal to that of Y.

Upper bound: Activate this option if you want to add lower bounds on the graph. Then select in this field the data to be used as upper bounds. The number of columns (column mode) or rows (row mode) to be selected must be equal to that of Y.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selected data (X, Y, Z, Groups, Weights and observation labels) contains a label.

Charts tab:

Chart type: select the type of chart you want to display:

- Bar chart.
- Curve.
- Scatter plot.

Example

A tutorial on how to a chart with error bars is available on XLSTAT Help Center:

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

Word cloud

Use this method to create a visual representation of text data, typically to represent keywords based on their frequencies. Tags are usually single words, and the frequency of each tag is shown by varying font size and/or color.

This format is useful for quickly identifying the most important terms in a document.

In this section:

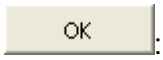
Description

Word cloud is a visualization tool based on frequency. In its simplest form, only one dimension of information is shown: font size is proportional to the word frequency, which means that the larger a word is in the word cloud, the more frequent the word is in the document.

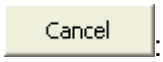
The word cloud feature takes as input a vector of terms (term labels) and a term frequency matrix (one column per document). Using that information, it plots one word cloud per document.

A custom color scale can be specified in order to customize the word colors, depending on their relative frequency.

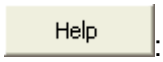
Dialog box



: Click this button to create the charts.



: Click this button to close the dialog box without creating the charts.



: Click this button to display help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables

(column mode). If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations (row mode).

General tab:

Term frequency matrix: Select in this field the term frequency matrix. If a column header has been selected, check that the "Column/Row labels" option has been activated.

Term labels: Select in this field the term labels data. If a column header has been selected, check that the "Column/Row" option has been activated.

Colory by: * **Frequencies**: Choose this option so that the words color depends on their frequency. * **Values**: Choose this option so that the words color depends on a value that you provide. In that case you must select the corresponding data. If the "column/row labels" option is checked, make sure you select a header as well for this field.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row/column of the selected data (term frequency matrix, term labels, values) contains a header.

Options tab:

Max. words: Activate this option to set a maximum number of words to be plotted in the word cloud. Least frequent terms will be dropped.

Random position: Activate this option to plot words in random order. If not, they will be plotted in decreasing frequency from the center of the plot.

Rotation period: Activate this option to set the rotation period of 90° between each displayed word. By default the rotation period is 4 words, i.e. one rotation at 90 degree will be applied to the cloud display sequence every fourth word.

Color scale: Select the type of color scale.

- **Preset**: Activate this option to choose a preset color scale. Activate the **Log scale** option if you want to emphasize the color difference among the low frequency terms.
- **Random color scale**: Activate this option so that a four colors scale is randomly generated.
- **Custom color scale**: Activate this option to manually select the cells corresponding to the color scale. The background color of the cells is used to obtain the color that will be used for the words. Colors must be ordered in decreasing frequency.

Results

The result that is displayed is one or several Word Cloud charts (depending on the number of columns in the Term frequency matrix). As it is an Excel chart, you can modify colors, font type, words position as much as you want.

Example

An example of a Word Cloud generated with XLSTAT is available at:

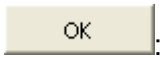
http://www.xlstat.com/demo-wdc.htm

Radar charts

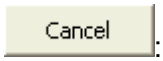
Use this tool to create radar charts (also called spider charts, or star charts). This type of chart is useful for visualizing and comparing three or more quantitative variables. For example, we can use a radar chart to compare students' grades in each course to the class average. XLSTAT allows you to display the labels around the chart either in the form of text or in the form of images.

In this section:

Dialog box



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Variables or **Images**: XLSTAT allows you to select variable names in two forms: as text or as images. Select in this field the labels or images, depending on the option chosen, that will appear around the chart.

Data: Select the values of each variable.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selected data contains a label.

Options tab:

Chart type: XLSTAT allows you to customize your radar chart: * **Classic**: Enable this option to display a classic radar chart.

- **With points**: Activate this option to display a radar chart with points.
- **Filled**: Activate this option to color the different shapes of the radar chart.
- **Polar Graph**: Activate this option to display a polar chart.

Group variables: Activate this option to group variables on the same chart.

Labels: Enable this option to display labels on the radar chart.

Axis Values: Enable this option to display the axis values on the chart.

Example

A tutorial on how to use radar charts is available on XLSTAT Help Center:

http://www.xlstat.com/radar_en.htm

Tornado diagrams

Use this tool to create a tornado chart or a back-to-back histogram from a sample of continuous or discrete quantitative data.

In this section:

Description

Tornado diagrams

The tornado chart is a visualization tool similar to a bar chart that allows you to compare the relative importance of two variables. Categories are in general ordered so that the largest bar appears at the top of the graph, the second largest at the second and so on, but XLSTAT allows you to unorder the categories if you wish. In the end, the graphic looks like a tornado, hence its name.

Histograms

The histogram is one of the most frequently used visualization tools as it gives a very quick idea of the distribution of a sample of continuous or discrete data.

Intervals definition

One of the challenges in creating histograms is to define the intervals, as for a determined set of data, the shape of the histogram depends solely on the definition of the classes. Between the two extremes of the single class comprising all the data and giving a single bar and the histogram with one value per class, there are as many possible histograms as there are data partitions.

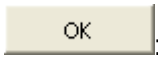
To obtain a visually and operationally satisfying result, defining classes may require several attempts.

The most traditional method consists of using classes defined by intervals of the same width, the lower bound of the first interval being determined by the minimum value or a value slightly less than the minimum value.

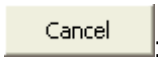
To make it easier to obtain histograms, XLSTAT lets you create histograms either by defining the number of intervals, their width or by specifying the intervals yourself. The intervals are considered as closed for the lower bound and open for the upper bound.

Dialog box

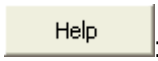
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to samples. If the arrow points to the right, XLSTAT considers that rows correspond to samples and columns to observations.


: Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Format:

Histograms: Choose this option if you want XLSTAT to display a back-to-back histogram.

Tornado diagrams: Choose this option if you want XLSTAT to display a tornado diagram.

Data: Select two columns of quantitative data or one column of quantitative data and a column with two groups. If headers have been selected, please ensure that the "Column Labels" option is enabled.

If you choose to display a back-to-back histogram:

Data type:

Continuous: Choose this option so that XLSTAT considers your data to be continuous.

Discrete: Choose this option so that XLSTAT considers your data to be discrete.

If you choose to display a tornado diagram:

Labels: Select this option if you want to enter a column of labels.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selections contains a label.

Groups: Activate this option then select a column (column mode) or a row (row mode) containing the data descriptors. If a header was selected, please verify that the "Column Labels" option is enabled.

Options tab:

Chart title: Enter the chart title you want to display on the bar chart.

If you choose to display a tornado diagram:

Bars: Choose whether you want to display **vertical** or **horizontal** bars.

Sort: Choose this option if you want to sort the tornado chart data.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics of the samples.

Results

Summary statistics: This table displays for the selected samples, the number of observations, the number of missing values, the number of non- missing values, the mean and the standard deviation.

Tornado diagrams: Tornado diagrams are displayed. If you wish, you can change line color, scales, and titles as with any Excel chart.

Histograms: The histograms are displayed. If desired, you can change the color of the lines, scales, titles as with any Excel chart.

Descriptive statistics for the intervals: This table displays for each interval its lower bound, upper bound, the frequency (number of values of the sample within the interval), the relative frequency (the number of values divided by the total number of values in the sample), and the density (the ratio of the frequency to the size of the interval).

Example

An example showing how to create a tornado diagram is available on XLSTAT Help Center:

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

Funnel Charts

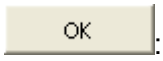
Use this tool to generate funnel charts, ideal for visualizing how data is filtered through stages of a process or project. This chart looks like a funnel, where each segment gradually narrows. In a funnel chart, the width of a bar is proportional to the maximum value of the measure. The top bar is always equal to 100% and the size of the following bars is defined relative to that of the first. It quickly allows you to see at which stages a decline occurs and at what rate.

This type of chart is primarily used in marketing or human resources with the goal of: * Showing the number of prospects at each stage of the sales process * Visualizing a buyer's journey * Seeing the number of respondents to a survey or recruitment process

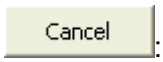
It is therefore very practical for illustrating the different stages of a process and the overall reduction of each of them compared to the previous one.

In this section:

Dialog box



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

![Select_mouse.png][image-8]{: width="26" height="25"} ![Select_list.png][image-9]{: width="26" height="25"} ![Select_file.png][image-10]{: width="26" height="25"} : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files ![Select_file_choosefile.png][image-11]{: width="26" height="25"}.

General tab:

Descriptors: Select the labels of the descriptors representing the stages of your process.

Values: Select the numerical data you want to represent on the graph.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet and then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Labels included: Check this option if the first line of the selected data contains a label.

Options tab:

Chart Title: Enter the title of the chart to create.

Percentages: Check this option to display percentages relative to the largest value.

- **Colors**:
  - **Distinct colors**: Each group is displayed with a distinct color.
  - **Color scale**: A gradual color scale is used so that the group with the greatest value is displayed in dark. It is possible to choose from different scale colors.

Example

A tutorial on how to use funnel chart is available on XLSTAT Help Center:

https://help.xlstat.com/6853-courbes-de-niveau-dans-excel-avec-xlstat

Contour plot

Use this tool to generate a contour plot or a surface plot. This type of visualization enables you to explore the relationship between a response variable and two predictor variables.

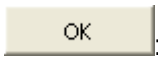
A contour plot represents a three-dimensional surface by drawing contours on a two-dimensional plane. The X and Y values are plotted along the respective axes, while the contour lines and bands depict the Z value.

This type of chart finds extensive application in cartography, where contour lines signify consistent elevations.

To ensure an accurate representation, the data matrix comprising the three variables X, Y, and Z must be well-balanced. Should it be otherwise, XLSTAT employs bilinear interpolation to fill in the missing values. Additionally, bicubic interpolation is also an option. The choice between these methods depends on the specifics of your dataset, as each may yield superior results under different circumstances.

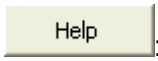
In this section:

Dialog box



: Click this button to start the calculations.

: Click this button to close the dialog box without doing any calculations.



: Click this button to display help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

![Select_mouse.png][image-8]{: width="26" height="25"} ![Select_list.png][image-9]{: width="26" height="25"} ![Select_file.png][image-10]{: width="26" height="25"} : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files ! [Select_file_choosefile.png][image-11]{: width="26" height="25"}.

General tab:

X: Select the numeric data for the first prediction variable X in this field.

Y: Select the numerical data for the second prediction variable Y in this field.

Y: Select the numeric data for the response variable Z in this field.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Labels included: Check this option if the first line of the selected data contains a label.

Options tab:

Chart type: XLSTAT allows you to display charts in different formats: * **Contour plot**: Enable this option to display a contour plot.

- **Surface plot**: Enable this option to display a surface plot.
- **Surface plot with grid**: Enable this option to display a surface plot with a grid.

3D View: Enable this option to view graphics in 3D.

Interpolation: XLSTAT allows you to perform data interpolation: * **Bilinear interpolation**: enable this option to perform bilinear interpolation of the data. This is particularly useful when the data is not balanced.

- **Bicubic interpolation**: enable this option to perform a bicubic interpolation of the data. This is particularly useful when the data is not balanced.

Example

A tutorial on how to use contour plots is available on XLSTAT Help Center:

https://help.xlstat.com/6853-courbes-de-niveau-dans-excel-avec-xlstat

Bar charts

Use this tool to quickly display bar charts which gives the possibility to use either classic labels or images as labels and/or as bar backgrounds. If your data correspond to countries you can automatically use the XLSTAT flags library. You can also input your own images.

In this section:

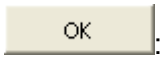
Description

This tool has been developed to help you quickly build charts that are more self-explanatory.

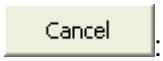
Should you have requests for more options, let us know, we will be happy to help you save more time.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Labels: Select the data that describe the values, and that might be used as labels on the charts.

- **Countries (Names)**: Choose this option if the *labels* correspond to the names of the countries.
- **Countries (Codes)**: Choose this option if the *labels* correspond to the (ISO 3166) two-letter codes of the countries.
- **Other**: Choose this option if the *labels* do not correspond to countries. This option requires that you have a set of images that you can select in the *Images* field.

Values: Select the data that are used on the chart for the bars.

Images: Select the the cells that contain for each label the images that you want to use.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row/column of the data selections includes a header.

Charts size:

- **Width**: Enter the value in points of the chart's width.
- **Height**: Enter the value in points of the chart's height.

Options tab:

Chart title: Enter the chart title you want to display on the bar chart.

Bars: Choose whether you want to display **vertical** or **horizontal** bars.

Labels: Choose whether you want to display the labels names or codes (only active if you selected countries as labels) on the charts. You can choose not to display either of the two.

Images position: You can display images on the side of the axis and/or in the bars themselves.

- **Next to axis**: Check this option if you want to display the images next to the axis. If your labels correspond to countries, you choose to display the flags in circles, squares or rectangles.
- **Images in bars**: Check this option if you want to display the images in the bars themselves. You can either display stretched or tiled images.

Results

XLSTAT displays one bar chart for each series you have selected in the values field.

Example

An example showing how to display bar charts that use images is available on the XLSTAT Help Center at

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

Truncated bar charts

Use this tool to create bar charts where part of the scale is squeezed.

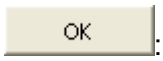
In this section:

Description

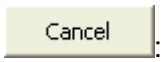
This visualization tool allows to display properly values that are spread on two extremes of the same scale, without having to use a log scale. Although it is very convenient, it is very cumbersome to create such a chart in Excel. Furthermore, XLSTAT allows to use transparency to partly show the information that is hidden in the removed or squeezed part of the scale.

This tool can work directly on quantitative data (typically frequencies), or on raw qualitative data (weighted or not) that are transformed into a frequencies table before the bar chart is displayed.

Dialog box



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data format: Select the data format.

- **Quantitative**: Activate this option if your data correspond to a simple list of numerical values, a contingency table or a crosstab. If the "Labels included" option is activated, you must select both the header (column) and the category (row) labels together with the data.

- **Qualitative**: Activate this option if your data are available as qualitative variables that need to be transformed to frequencies before the chart is created.

Quantitative data: If the data format is "quantitative", select the data that correspond to the quantitative data that must be displayed on the bar chart.

- **One chart**: Activate this option if you want the results corresponding to each column of quantitative data to be displayed on a single chart.

Qualitative data: If the data format is "qualitative", select the data that correspond to the qualitative data that will be used to construct a frequency table, that will then be displayed on the bar chart.

- **Sort the categories alphabetically**: Activate this option to sort alphabetically the categories of the qualitative variables. If this option is unchecked, the order of appearance is respected.

Subsamples: Check this option to select a column showing the names or indexes of the sub-samples for each of the observations.

- **Sort the categories alphabetically**: Activate this option to sort alphabetically the categories of the subsample data. If this option is unchecked, the order of appearance is respected.

- **Variable-Category labels**: Activate this option to use variable-category labels when displaying outputs. Variable-Category labels include the variable name as a prefix and the category name as a suffix.

- **One chart**: Activate this option if you want the results corresponding to each sub-sample to be displayed on a single chart.

Weights: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Labels included: Activate this option if the row and column labels of the quantitative data are selected.

Variable labels: Activate this option if the first row of the data selections includes a header.

Display title: Check this option to display a title on the chart.

Truncate: Check this option to squeeze the vertical axis of the chart (or horirontal axis if the horizontal option is selected). Select the start and end points on the scale that you want to squeeze or hide.

Transparency(%): Set the transparency of the squeezed parf of the chart. Set 0 to completely hide it, or 100 to make it fully visible.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

Example

A tutorial on generating a truncated bar chart is available at:

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

Plot a function

Use this tool to create a chart and plot a function on it, or to add a function to an existing chart.

In this section:

Description

This tool allows you to plot a function of the type $y = f(x)$ on an existing or new chart. The syntax of the function must respect the conventions imposed by Excel for functions used in spreadsheets. In addition, the abscissa must be identified by X1.

Examples:

Function	XLSTAT Syntax
$Y = x^2$	X1²
$Y = ln(x)$	LN(X1)
$Y = e(x)$	EXP(X1)
$Y =\mid x \mid$	ABS(X1)
$Y = x$ si $x < 0$, $Y = 2x$ si $x = 0$	IF(X1<0; X1; 2*X1)

In addition, you can as well use XLSTAT worksheet functions. For example, to plot the normal cumulative distribution function, enter XLSTAT_CDFNormal (X1).

Dialog box

OK: Click this button to create the chart.

Cancel: Click this button to close the dialog box without creating the chart.

Help: Click this button to display help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

General tab:

Function Y =: Enter the function that you want to plot, while respecting the syntax defined in the Description section.

Minimum: Enter the minimum value for which the function must be evaluated and plotted.

Max imum: Enter the maximum value for which the function must be evaluated and plotted.

Number of points: Enter the number of points at which the function must be evaluated between the minimum and maximum values. This option allows you to adjust the quality of the graph. For a function with many inflection points, too few points might give a graph of poor quality. Too many points may also degrade the quality of the display.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Active chart: Activate this option to add the function to the chart that is currently being selected.

Example

An example showing how to create a chart with a function is available on XLSTAT Help Center:

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

AxesZoomer

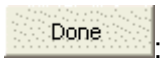
Use this tool to change the minimum and maximum values on the X- and Y-axes of a plot.

In this section:

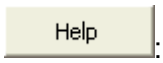
Dialog box

Important: before running this tool, you must select a scatter plot or curve.

: Click this button to apply changes to the plot.



: Click on this button to close the dialog box.



: Click this button to display help.

Min X: Enter the minimum value of the X-axis.

Max X: Enter the maximum value of the X-axis.

Min Y: Enter the maximum value of the X-axis.

Max Y: Enter the maximum value of the Y-axis.

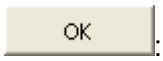
EasyLabels

Use this tool to add labels, formatted if required, to a series of values on a chart.

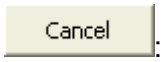
In this section:

Dialog box

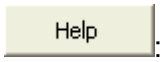
Important: before running this tool, you must select a scatter plot or curve, or a series of points on a plot.



: Click this button to add the labels.



: Click this button to close the dialog box without making any changes.



: Click this button to display help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that the labels are in a column. If the arrow points to the right, XLSTAT considers that the labels are in a row.

Labels: Select the labels to be added to the series of values selected on the plot.

Header in the first cell: Check this option if the first cell of the labels selected is a header and not a label.

Use the text properties: Check this option if you want the text format used in the cells containing the labels to also be applied to the text of labels in the chart:

- **Font**: Check this option to use the same character font.
- **Size**: Check this option to use the same size of font.
- **Style**: Check this option to use the same font style (normal, bold, italic).
- **Color**: Check this option to use the same font color.

Use the cell properties: Check this option if you want the format applied to the cells containing the labels to also be applied to the labels in the chart:

- **Border**: Check this option to use the same border.
- **Pattern**: Check this option to use the same pattern.

Use the point properties: Check this option if you want the label color to be the same as the color of the points:

- **Inside color**: Check this option to use the color inside the points.
- **Border color**: Check this option to use the border color of the points.

Reposition labels

Use this tool to change the position of observation labels on a chart.

In this section:

Dialog box

OK: Click this button to reposition the labels.

Cancel: Click this button to close the dialog box without making any changes.

Help: Click this button to display help.

 : Click this button to reload the default options.

Corners: Check this option to place labels in the direction of the corner of the quadrant in which the point is located.

Distance to point:

- **Automatic**: Check this option for XLSTAT to automatically determine the most appropriate distance to the point.
- **User defined**: Check this option to enter the value (in pixels) of the distance between the label and the point.

Above: Check this option to place labels above the point.

Right: Check this option to place labels to the right of the point.

Below: Check this option to place labels below the point.

Left: Check this option to place labels to the left of the point.

Apply only to the selected series: Check this option to only change the position of labels for the series selected.

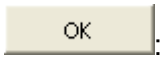
EasyPoints

Use this tool to modify the size, the color or the shape of the points that are displayed in an Excel chart.

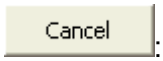
In this section:

Dialog box

Important: before running this tool, you must select a scatter plot or curve, or a series of points on a plot.



: Click this button to add the labels.



: Click this button to close the dialog box without making any changes.

: Click this button to display help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that the labels are in a column. If the arrow points to the right, XLSTAT considers that the labels are in a row.

Size: Activate this option and select the cells that give the size to be applied to the points. The size of the points is determined by the values in the cells.

Header in the first cell: Check this option if the first cell of the labels selected is a header and not a label.

Rescale: Choose the interval of sizes to use when displaying the points. The minimum must be between 2 and 71, and the maximum between 3 and 72.

Shapes and/or color: Activate this option to change the shape of the points and/or the color to be applied to the points. Select the cells and which color (if the Use the cell properties that tell which shape should be used for each point: 1 corresponds to a square, 2 to a diamond, 3 to a

triangle, 4 to an x, 5 to a star (*), 6 to a point (.), 7 to a dash (-), 8 to a plus (+) and 9 to a circle (o). The color of the border of the points depends on the color of the bottom border of the cells and the inside color of the points depends on the background color of the cells (Note: the default color of the cells is "none", so you need to set it to white to obtain white points).

Change shapes: Check this option if you want the shapes to be changed depending on the values selected in the "Shapes and or color" field.

Use the cell properties: Check this option if you want the format applied to the cells to also be applied to the points in the chart:

- **Border**: Check this option to use the cell borders as the foreground color.
- **Background**: Check this option to use the cell color as the background color.

Example

An example describing how to use the EasyPoints tool is available on the XLSTAT Help Center at:

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

Color, thickness and size

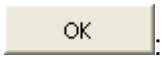
Use this tool to change the color or line thickness or point size of multiple series, in one click.

In this section:

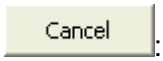
Description

This tool allows you to change the color and/or line thickness and point size of multiple series, in a single operation.

Dialog box



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

Colors: Select the cells containing the colors to apply to the lines or points. If your chart has S series, you must select S cells. Their background color will be applied to the series. If you want to color each of the P points of the S series, you must select a table with S rows and P columns.

Lines thickness: Select the thickness (in units of Excel points) of the series in cells. If the thickness is unique, you can enter it in the selection field.

Points size: Select the cells giving the size (in units of Excel points) of the points. If the size is unique, you can enter it in the selection field. If you want to size each of the P points of the S series, you must select a table with S rows and P columns

Header in the first cell: Check this option if the first cell of the selected data is a header.

Example

An example on how to use this tool is available on XLSTAT Help Center:

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

Orthonormal plots

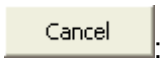
Use this tool to adjust the minimum and maximum of the X- and Y- axes so that the plot becomes orthonormal. This tool is particularly useful if you have enlarged an orthonormal plot produced by XLSTAT (for example after a PCA) and you want to ensure the plot is still orthonormal.

Note: an orthonormal plot is where a unit on the X-axis appears the same size as a unit on the Y-axis. Orthonormal plots avoid interpretation errors due to the effects of dilation or overwriting.

In this section:

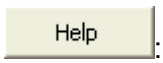
Dialog box

: Click this button to apply the transformation to the plot.



: Click this button to cancel the transformation of the plot.

: Click on this button to close the dialog box.



: Click this button to display help.

Resize a chart

Use this tool to resize a chart, or a plot area inside a chart.

In this section:

Dialog box

: Click this button to resize the chart.

: Click on this button to close the dialog box.

: Click this button to display help.

Select the zone you want to resize:

- **Chart**: Activate this option to resize the whole chart.
- **Plot area**: Activate this option to resize only the plot area.

Original size: The width and the height displayed here correspond to those of the selected chart or plot area before the resizing.

New size: Enter the new width and the new height of the chart, either in pixels or in percentage of the original size.

Lock aspect ratio: Activate this option if you want to that the initial proportions of the chart are respected.

Plot transformations

Use this tool to apply one or more transformations to the points in a plot.

In this section:

Dialog box

Important: before running this tool, you must select a scatter plot or curve.

OK: Click this button to transform the plot.

Cancel: Click this button to close the dialog box without carrying out the transformation.

Help: Click this button to display help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

Symmetry:

- **Horizontal axis**: Check this option to apply a symmetry around the X-axis.
- **Vertical axis**: Check this option to apply a symmetry around the Y-axis.

Note: if you select both the previous options, the symmetry applied will be a **central symmetry.**

Translation:

- **Horizontal**: Check this option to enter the number of units for a horizontal translation.
- **Vertical**: Check this option to enter the number of units for a vertical translation.

Rotation:

- **Angle (°)**: enter the angle in degrees for the rotation to be applied.
- **Right**: if this option is activated, a clockwise rotation is applied.

- **Left**: if this option is activated, an anti-clockwise rotation is applied.

Rescaling:

- **Factor**: enter the scaling factor to be applied to the data.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Display the new coordinates: Check this option to display the coordinates once all the transformations have been applied.

Update Min and Max on the new plot: Check this option for XLSTAT to automatically adjust the minimum and maximum of the X- and Y- axes, once the transformations have been carried out, so that all points are visible.

Orthonormal plot: Check this option for XLSTAT to automatically adjust the minimum and maximum of the X- and Y- axes, once the transformations have been carried out, so that the plot becomes orthonormal.

Merge plots

Use this tool to merge multiple plots into one.

In this section:

Dialog box

Important: before using this tool, you must select at least two plots of the same type (e.g. two scatter plots).

OK: Click this button to merge the plots.

Cancel: Click on this button to close the dialog box.

Help: Click this button to display help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

Display title: Check this option, to display a title on the merged plot.

- **Title of the first chart**: Check this option to use the title of the first chart.
- **New title**: Check this option to enter a title for the merged plot.

Orthonormal plot: Check this option for XLSTAT to verify that the plot resulting from the merged plots is orthonormal.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet in the active workbook.

Workbook: Activate this option to display the results in a new workbook.

New chart sheet: Check this option to display the plot resulting from the merge in a new chart sheet.

Display the report header: clear this option to stop the previous report header for the chart from being displayed.

Analyzing data

Factor analysis

Factor analysis highlights, where possible, the existence of underlying factors common to the quantitative variables measured in a set of observations.

In this section:

Description

The factor analysis method dates from the start of the 20th century (Spearman, 1904) and has undergone a number of developments, several calculation methods having been put forward. This method was initially used by psychometricians, but its field of application has little by little spread into many other areas, for example, geology, medicine and finance.

Today, there are two main types of factor analysis:

- Exploratory factor analysis (or EFA)
- Confirmatory factor analysis (or CFA)

It is EFA which will be described below and which is used by XLSTAT. It is a method which reveals the possible existence of underlying factors which give an overview of the information contained in a very large number of measured variables. The structure linking factors to variables is initially unknown and only the number of factors may be assumed.

CFA in its traditional guise uses a method identical to EFA but the structure linking underlying factors to measured variables is assumed to be known. A more recent version of CFA is linked to models of structural equations.

Going from p variables to k factors

Spearman's historical example, even if the subject of numerous criticisms and improvements, may still be used to understand the principle and use of the method. By analyzing correlations between scores obtained by children in different subjects, Spearman wanted to form a

hypothesis that the scores depended ultimately on one factor, intelligence, with a residual part due to an individual, cultural or other effect.

Thus the score obtained by an individual (i) in subject (j) could be written as $x(i,j) = \mu + b(j)F + e(i,j)$, where μ is the average score in the sample studied and F the individual's level of intelligence (the underlying factor) and e(i,j) the residual.

Generalizing this structure to p subjects (the input variables) and to k underlying factors, we obtain the following model:

$$x = \mu + \Lambda f + u \tag{1}$$

where x is a vector of dimension (p x 1), μ in the mean vector, Λ is the matrix (p x k) of the factor loadings and f and u are the random vectors of dimensions (k x 1) and (p x 1) respectively are assumed to be independent. The elements of f are called common factors, and those of u specific factors.

If we set the norm of f to 1, then the covariance matrix for the input variables from expression (1) is written as:

$$\Sigma = \Lambda\Lambda' + \Psi \tag{2}$$

Thus the variance of each of the variables can be divided into two parts: The communality (as it arises from the common factors),

$$h_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2 \tag{3}$$

and Ψ_{ii} the specific variance or unique variance (as it is specific to the variable in question).

It can be shown that the method used to calculate matrix Λ, an essential challenge in factorial analysis, is independent of the scale. It is therefore equivalent to working from the covariance matrix or correlation matrix.

The challenge of factorial analysis is to find matrices Λ and Ψ, such that equation (2) can be at least approximately verified.

Note: factor analysis is sometimes included with Principle Component Analysis (PCA) as PCA is a special case of factor analysis (where k, the number of factors, equals p, the number of variables). Nevertheless, these two methods are not generally used in the same context. Indeed, PCA is first and foremost used to reduce the number of dimensions while maximizing the unchanged variability in order to obtain independent (non-correlated) factors or for visualizing data in a 2- or 3-dimensional space. Whereas, factor analysis is used to identify a latent structure and for possibly reducing afterwards the number of variables measured if they are redundant with respect to the latent factors.

Extracting Factors

Three methods of extracting latent factors are offered by XLSTAT:

Principle components: this method is also used in Principle Component Analysis (PCA). It is only offered here in order to make a comparison between the results of the three methods bearing in mind that the results from the module dedicated to PCA are more complete.

Principal factors: this method is probably the most used. It is an iterative method which enables the communalities to be gradually converged. The calculations are stopped when the maximum change in the communalities is below a given threshold or when a maximum number of iterations is reached. The initial communalities can be calculated according to various methods.

Maximum likelihood: this method was first put forward by Lawley (1940). The proposal to use the Newton-Raphson algorithm (iterative method) dates from Jennrich (1969). It was afterwards improved and generalized by Jöreskog (1977). This method assumes that the input variables follow a normal distribution. The initial communalities are calculated according to the method proposed by Jöreskog (1977). As part of this method, an adjustment test is calculated. The statistic used for the test follows a Chi-square distribution to $(p-k)^2/2-(p+k)/2$ degrees of freedom where p is the number of variables and k the number of factors.

Number of factors

Determining the number of factors to select is one of the challenges of factor analysis. The "automatic" method offered by XLSTAT is uniquely based on the spectral decomposition of the correlation matrix and the detection of a threshold from which the contribution made by information (in the sense of variability) is not significant.

The likelihood maximum method offers an adjustment test to help determine the correct number of principle factors for the principle factor method. For the principal factors method, the defining the number of factors is more difficult?

The Kaiser-Guttman rule suggests that only those factors with associated eigenvalues which are strictly greater than 1 should be kept. The number of factors to be kept corresponds to the first turning point found on the curve. Crossed validation methods have been suggested to achieve this aim.

Anomalies (Heywood cases)

Communalities are by definition the squares of correlations. They must therefore be between 0 and 1. However, it may happen that the iterative algorithms (principle factors method or likelihood maximum method) will produce solutions with communalities equal to 1 (Heywood cases), or greater than 1 (ultra Heywood cases). There may be many reasons for these anomalies (too many factors, not enough factors, etc.). When this happens, XLSTAT sets the communalities to 1 and adapts the elements of Λ in consequence.

Rotations

Once the results have been obtained, they may be transformed in order to make them more easy to interpret, for example by trying to arrange that the coordinates of the variables on the factors are either high (in absolute value), or near to zero. There are two main families of rotations:

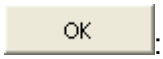
Orthogonal rotations can be used when the factors are not correlated (hence orthogonal). The methods offered by XLSTAT are Varimax, Quartimax, Equamax, Parsimax and Orthomax. Varimax rotation is the most used. It ensures that for each factor there are few high factor loadings and few that are low. Interpretation is thus made easier as, in principle, the initial variables will mostly be associated with one of the factors.

Oblique transformations can be used when the factors are correlated (hence oblique). The methods offered by XLSTAT are Quartimin and Oblimin.

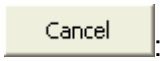
The Promax method, also offered by XLSTAT, is a mixed procedure since it consists initially of a Varimax rotation followed by an oblique rotation so that the high factor loadings and low factor loadings are the same but with the low values even lower.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

The main data entry field is used to select one of three types of table:

Observations/variables table / Correlation matrix / Covariance matrix: Choose the option appropriate to the format of your data, and then select the data. If your data correspond to a

table comprising N observations described by P quantitative variables select the Observations/variables option. If column headers have been selected, check that the "Variable labels" option has been activated. If you select a correlation or covariance matrix, and if you include the variable names in the first row of the selection, you must also select them in the first column.

Correlation: Choose the type of matrix to be used by factor analysis. The difference between the Pearson (n) and the Pearson (n-1) options, only influences the way the variables are standardized, and the difference can only be noticed on the coordinates of the observations.

Extraction method: Choose the factor extraction method to be used, The three possible methods are (see the description section for more details):

- Principal components
- Principal factors
- Maximum likelihood

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (input table, weights, observation labels) includes a header. Where the selection is a correlation or covariance matrix, if this option is activated, the first column must also include the variable labels.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Number of factors:

- **Automatic**: Activate this option to make XLSTAT determine the number of factors automatically.
- **User defined** Activate this option to tell XLSTAT the number of factors to use in the calculations.

Initial communalities: Choose this calculation method for the initial communalities (this option is only visible for the principle factors methods):

- **Squared multiple correlations**: The initial communalities are based a variable's level of dependence with regard to the other variables.
- **Random**: The initial communalities are drawn from the interval]0 ; 1[.
- **1**: The initial communalities are set to 1.
- **Maximum**: The initial communalities are set to the maximum value of the squares of the multiple correlations.

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 50.
- **Convergence**: Enter the maximum value of the evolution in the communalities from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.0001.

Rotation: Activate this option if you want to apply a rotation to the factor coordinate matrix.

- **Number of factors**: Enter the number of factors the rotation is to be applied to.
- **Method**: Choose the rotation method to be used. For certain methods a parameter must be entered (Gamma for Orthomax, Tau for Oblimin, and the power for Promax).
- **Kaiser normalization**: Activate this option to apply Kaiser normalization during the rotation calculation.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Pairwise deletion: Activate this option to remove observations with missing data only when the variables involved in the calculations have missing data. For example, when calculating the correlation between two variables, an observation will only be ignored if the data corresponding to one of the two variables is missing.

Estimate missing data: Activate this option to estimate the missing data before the calculation starts.

- **Mean or mode**: Activate this option to estimate the missing data by using the mean (quantitative variables) or the mode (qualitative variables) for the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data for an observation by searching for the nearest neighbor to the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

Correlations: Activate this option to display the correlation or covariance matrix depending on the type of options chosen in the "General" tab. If the **Test significance** option is activated, the significant correlations at the selected **significance threshold** are displayed in bold.

Kaiser-Meyer-Olkin: Activate this option to compute the Kaiser-Meyer- Olkin Measure of Sampling Adequacy.

Cronbach's Alpha: Activate this option to compute the Cronbach's alpha coefficient.

Eigenvalues: Activate this option to display the table and chart (scree plot) of eigenvalues.

Factor pattern: Activate this option to display factor loadings (coordinates of variables in the factor space).

Factor/Variable correlations: Activate this option to display correlations between factors and variables.

Factor pattern coefficients: Activate this option if you want the coefficients of the factor pattern to be displayed. Multiplying the (standardized) coordinates of the observations in the initial space by these coefficients gives the coordinates of the observations in the factor space.

Factor structure: Activate this option to display correlations between factors and variables after rotation.

Charts tab:

Variables charts: Activate this option to display charts representing the variables in the new space.

- **Vectors**: Activate this option to display the initial variables in the form of vectors.

Correlations charts: Activate this option to display charts showing the correlations between the factors and initial variables.

- **Vectors**: Activate this option to display the initial variables in the form of vectors.

Observations charts: Activate this option to display charts representing the observations in the new space.

- **Labels**: Activate this option to have observation labels displayed on the charts. The number of labels displayed can be changed using the filtering option.

Colored labels: Activate this option to show labels in the same color as the points.

Filter: Activate this option to modulate the number of observations displayed:

- **Random**: The observations to display are randomly selected. The "Number of observations" N to display must then be specified.
- **N first rows**: The N first observations are displayed on the chart. The "Number of observations" N to display must then be specified.
- **N last rows**: The N last observations are displayed on the chart. The "Number of observations" N to display must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to display.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. This includes the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

Correlation/Covariance matrix: This table shows the data to be used afterwards in the calculations. The type of correlation depends on the option chosen in the "General" tab in the dialog box. For correlations, significant correlations are displayed in bold.

Measure of Sample Adequacy of Kaiser-Meyer-Olkin: This table gives the value of the KMO measure for each individual variable and the overall KMO measure. The KMO measure ranges between 0 and 1. A low value corresponds to the case where it is not possible to extract synthetic factors (or latent variables). In other words, observations do not bring out the model that one could imagine (the sample is "inadequate"). Kaiser (1974) recommends not to accept a factor model if the KMO is less than 0.5. If the KMO is between 0.5 and 0.7 then the quality of the sample is mediocre, it is good for a KMO between 0.7 and 0.8, very good between 0.8 and 0.9 and excellent beyond.

Cronbach's Alpha: If this option has been activated, the value of Cronbach's Alpha is displayed.

Maximum change in communality at each iteration: This table is used to observe the maximum change in communality for the last 10 iterations. For the maximum likelihood method, the evolution of a criterion which is proportional to the opposite of the likelihood maximum is also displayed.

Goodness of fit test: The goodness of fit test is only displayed when the likelihood maximum method has been chosen.

Reproduced correlation matrix: This matrix is the product of the factor loadings matrix with its transpose.

Residual correlation matrix: This matrix is calculated as the difference between the variables correlation matrix and the reproduced correlation matrix.

Eigenvalues: This table shows the eigenvalues associated with the various factors together with the corresponding percentages and cumulative percentages.

Eigenvectors: This table shows the eigenvectors.

Factor pattern: This table shows the factor loadings (coordinates of variables in the vector space, also called *factor pattern*). The corresponding chart is displayed.

Factor/Variable correlations: This table displays the correlations between factors and variables.

Factor pattern coefficients: This table displays the coefficients of the factor pattern to be displayed. Multiplying the (standardized) coordinates of the observations in the initial space by these coefficients gives the coordinates of the observations in the factor space.

Where a rotation has been requested, the results of the rotation are displayed with the **rotation matrix** first applied to the factor loadings. This is followed by the modified variability percentages associated with each of the axes involved in the rotation. The coordinates of the variables and observations after rotation are displayed in the following tables.

Factor structure: This table shows the correlations between factors and variables after rotation.

Example

A tutorial on how to use Factor analysis is available on XLSTAT Help Center:

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

References

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.

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.

Jöreskog 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 & Sons, New York.

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.

Principal Component Analysis (PCA)

Use Principle Component Analysis to analyze a quantitative observations/variables table or a correlation or covariance matrix. This method is used to:

- Study and visualize correlations between variables.
- Obtain non-correlated factors which are linear combinations of the initial variables.
- Visualize observations in a 2- or 3-dimensional space.

In this section:

Description

Principle Component Analysis (PCA) is one of the most frequently used multivariate data analysis methods. Given a table of quantitative data (continuous or discrete) in which n observations (observations, products, etc.) are described by p variables (the descriptors, attributes, measurements, etc.), if p is quite high, it is impossible to grasp the structure of the data and the nearness of the observations by merely using univariate statistical analysis methods or even a correlation matrix.

Uses of PCA

There are several uses for PCA, including:

The study and visualization of the correlations between variables to hopefully be able to limit the number of variables to be measured afterwards;

Obtaining non-correlated factors which are linear combinations of the initial variables so as to use these factors in modeling methods such as linear regression, logistic regression or discriminant analysis.

Visualizing observations in a 2- or 3-dimensional space in order to identify uniform or atypical groups of observations.

Principle of PCA

PCA can be considered as a projection method which projects observations from a p-dimensional space with p variables to a k-dimensional space (where $k < p$) so as to conserve the maximum amount of information (information is measured here through the total variance of the scatter plots) from the initial dimensions. If the information associated with the first 2 or 3 axes represents a sufficient percentage of the total variability of the scatter plot, the observations will be able to be represented on a 2- 3-dimensional chart, thus making interpretation much easier.

Correlations or covariance

PCA is used to calculate matrices to project the variables in a new space using a new matrix which shows the degree of similarity between the variables. It is common to use the Pearson correlation coefficient or the covariance as the index of similarity, Pearson correlation and covariance have the advantage of giving positive semi-defined matrices whose properties are used in PCA. However other indexes may be used. XLSTAT makes it possible to also use the Spearman correlation coefficient because the corresponding matrices are also positive semi-definite. Making a PCA on a Spearman correlation matrix is fully equivalent to a classic PCA (based on Pearson correlation) performed on the matrix of ranks. When you run a PCA based on Spearman correlations, XLSTAT offers the option to display the matrix of the ranks in the report.

Traditionally, a correlation coefficient rather than the covariance is used as using a correlation coefficient removes the effect of scale: thus a variable which varies between 0 and 1 does not weigh more in the projection than a variable varying between 0 and 1000. However in certain areas, when the variables are supposed to be on an identical scale or we want the variance of the variables to influence factor building, covariance is used.

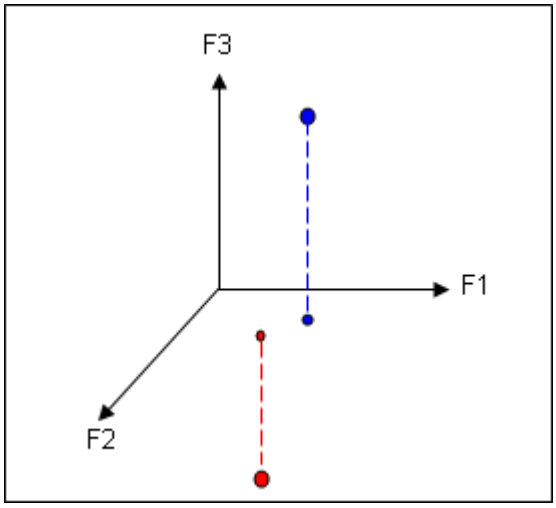
Where only a similarity matrix is available rather than a table of observations/variables, or where you want to use another similarity index, you can carry out a PCA starting from the similarity matrix. The results obtained will only concern the variables as no information on the observations was available.

Note: where PCA is carried out on a correlation matrix, it is called normalized PCA.

Interpreting the results

Representing the variables in a space of k factors enables the correlations between the variables and between the variables and factors to be visually interpreted with certain precautions.

Indeed if the observations or variables are being represented in the factor space, two points a long distance apart in a k-dimensional space may appear near in a 2-dimensional space depending on the direction used for the projection (see diagram below).



We can consider that the projection of a point on an axis, a plan or a 3-dimensional space is reliable if the sum of the squared cosines on the representation axis is near to 1. The squared cosines are displayed in the results given by XLSTAT in order to avoid any incorrect interpretation.

If the factors are afterwards to be used with other methods, it is useful to study the relative contribution (expressed as a percentage or a proportion) of the different variables in building each of the factor axes so as to make the results obtained afterwards easy to interpret. The contributions are displayed in the results given by XLSTAT.

Number of factors

Two methods are commonly used for determining the number of factors to be used for interpreting the results:

The *scree test* (Cattell, 1966) is based on the decreasing curve of eigenvalues. The number of factors to be kept corresponds to the first turning point found on the curve.

We can also use the cumulative variability percentage represented by the factor axes and decide to use only a certain percentage.

Graphic representations

One of the advantages of PCA is that it simultaneously provides the best view of both variables and observations with biplots combining both (see below). However, these representations are only reliable if the sum of the variability percentages associated with the axes of the representation space are sufficiently high. If this percentage is high (for example 80%), the representation can be considered as reliable. If the percentage is reliable, it is recommended to produce representations on several axis pairs in order to validate the interpretation made on the first two factor axes.

Biplots

After carrying out a PCA, it is possible to simultaneously represent both observations and variables in the factor space. The first work on this subject dates from Gabriel (1971). Gower (1996) and Legendre (1998) synthesized the previous work and extended this graphical representation technique to other methods. The term biplot is reserved for simultaneous representations which respect the fact that the projection of observations on variable vectors must be representative of the input data for the same variables. In other words, the projected points on the variable vectors must respect the order and the relative distances of the observations for that same variable, in the input data.

The simultaneous representation of observations and variables cannot be produced directly by taking the coordinates of the variables and observations in the factor space. A transformation is required in order to make the interpretation precise. Three methods using the graphic representation are available depending on the type of interpretation desired:

Correlation biplot: This type of biplot interprets the angles between the variables as these are directly linked to the correlations between the variables. The position of two observations projected onto a variable vector can be used to determine their relative level for this variable. The distance between the two observations is an approximation of the Mahalanobis distance in the k-dimensional factor space. Lastly, the projection of a variable vector in the representation space is an approximation of the standard deviation of the variable (the length of the vector in the k-dimensional factor space is equal to the standard deviation of the variable).

Distance biplot: A distance biplot is used to interpret the distances between the observations as these are an approximation of their Euclidean distance in the p-dimensional variable space. The position of two observations projected onto a variable vector can be used to determine their relative level for this variable. Lastly, the length of a variable vector in the representation space is representative of the variable's level of contribution to building this space (the length of the vector is the square root of the sum of the contributions).

Symmetric biplot: This biplot was proposed by Jobson (1992) and is half-way between the two previous biplots. If neither the angles nor the distances can be interpreted, this representation may be chosen as it is a compromise between the two.

XLSTAT lets you adjust the lengths of the variable vectors so as to improve the readability of the charts. However, if you use this option with a correlation biplot, the projection of a variable vector will no longer be an approximation of the standard deviation of the variable.

Bootstrap charts

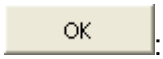
Several validation techniques have been developed for PCA. One objective of validation is to estimate the proximity between the observations on a factorial plan and to know which observations are significantly different from each other. For that purpose, XLSTAT uses the partial bootstrap method (Lebart, 2007). The partial bootstrap method consists in drawing m samples (with replacement) each of the same size as the matrix of data used for the PCA. Then, each sample is centered, and normalized in the case of normalized PCA, and each observation of each sample is displayed on the factorial plans as supplementary observation. As a consequence, a cloud of bootstrap observations is generated around each original observation. In order to simplify charts, XLSTAT proposes two types of representations:

- **Convex hulls**: XLSTAT looks for the bootstrap observations the most extreme and connects them to represent the convex hull of the cloud of points.
- **Confidence ellipses**: XLSTAT calculates and represents for each original observation the 95 % confidence ellipse based and centered on the bootstrap points.

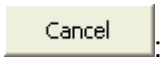
These two graphical representations can be interpreted in the same way. Indeed, we can conclude that two observations are significantly different from each other on a given factorial plan, if their convex hulls or ellipses do not overlap.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

The main data entry field is used to select one of three types of table:

Observations/variables table / Correlation matrix / Covariance matrix: Choose the option appropriate to the format of your data, and then select the data. If your data correspond to a table comprising n observations described by p quantitative variables, select the Observations/variables option. If column headers have been selected, check that the "Variable labels" option has been activated.

PCA type: If the format selected is "observations/variables", you can choose between **Correlation maxtrix** (standardized or normalized PCA), **Covariance maxtrix** (unstandardized or non-normalized PCA) and **Spearman**. Choose Spearman to perform PCA on a Spearman correlation matrix. If the format of the input data is **Covariance matrix**, you can choose

between correlation (the selected covariance matrix is transformed into a correlation matrix and a standardized PCA will be applied) or covariance, in which case, the PCA is performed on the selected covariance matrix.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (observations/variables table, weights, observation labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Filter factors: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %**: Activate this option then enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Activate this option to set the number of factors to take into account.

Standardisation: If the format of your data is "observations/variables", you can choose how correlation (or covariance) will be computed: with denominator (n) or (n – 1).

Rotation: Activate this option if you want to apply a rotation to the factor coordinate matrix.

- **Number of factors**: Enter the number of factors the rotation is to be applied to.
- **Method**: Choose the rotation method to be used. For certain methods a parameter must be entered (Kappa for Orthomax, Tau for Oblimin, and the power for Promax).
- **Kaiser normalization**: Activate this option to apply Kaiser normalization during the rotation calculation.

Supplementary data tab:

Supplementary observations: Activate this option if you want to calculate and represent the coordinates of additional observations. These observations are not taken into account for the computation of the correlation matrix and for the subsequent calculations (we talk of passive observations as opposed to active observations). If the first row of the data selection for supplementary observations includes a header you must activate the "Variable labels for supp. obs" option.

Supplementary variables: Activate this option if you want to calculate coordinates afterwards for variables which were not used in calculating the factor axes (passive variables as opposed to active variables).

- **Quantitative**: Activate this option if you have supplementary quantitative variables. If column headers were selected for the main table, ensure that a label is also present for the variables in this selection.
- **Qualitative**: Activate this option if you have supplementary qualitative variables. If column headers were selected for the main table, ensure that a label is also present for the variables in this selection.
- **Display the centroids**: Activate this option to display the centroids that correspond to the categories of the supplementary qualitative variables.

Data options tab:

Missing data:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Pairwise deletion: Activate this option to remove observations with missing data only when the variables involved in the calculations have missing data. For example, when calculating the correlation between two variables, an observation will only be ignored if the data corresponding to one of the two variables is missing.

Estimate missing data: Activate this option to estimate the missing data before the calculation starts.

- **Mean or mode**: Activate this option to estimate the missing data by using the mean (quantitative variables) or the mode (qualitative variables) for the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data for an observation by searching for the nearest neighbor to the observation.

Replace missing data by 0: if the format of your data is "correlation matrix" or "covariance matrix", you can activate this option to replace missing data by 0.

Groups:

By group analysis: Activate this option and select the data that describe to which group each observation belongs. You can choose between the following three options:

- **One PCA per group**: This option allows you to perform one PCA for each group.
- **One PCA per selected group**: This option allows you to choose in a new dialog box on which groups you want to run separate PCAs.
- **One PCA on merged groups**: This option allows you to perform a unique PCA on a series of groups, that will be merged once you have selected them in a dialog box that is displayed at the beginning of the computations.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the correlation or covariance matrix depending on the "PCA type" option chosen in the "General" tab.

- **Test significance**: Where a correlation was chosen in the "General" tab in the dialog box, activate this option to test the significance of the correlations.
- **Bartlett's sphericity test**: Activate this option to perform the Bartlett sphericity test.
- **Significance level (%)**: Enter the significance level for the above tests.
- **Kaiser-Meyer-Olkin**: Activate this option to compute the Kaiser-Meyer-Olkin Measure of Sampling Adequacy.

Matrix of ranks: If you have chosen to run the PCA on a Spearman correlatio matrix, you can display the matrix of ranks of your data with this option.

Eigenvalues: Activate this option to display the table and chart (scree plot) of eigenvalues.

Factor loadings: Activate this option to display the coordinates of the variables in the factor space.

Variables/Factors correlations: Activate this option to display correlations between factors and variables.

Factor scores: Activate to display the coordinates of the observations (factor scores) in the new space created by PCA.

Contributions: Activate this option to display the contribution tables for the active variables and observations.

Squared cosines: Activate this option to display the tables of squared cosines for the variables and observations.

Charts tab:

Variables sub-tab:

Correlations charts: Activate this option to display charts showing the correlations between the components and initial variables. This chart is named correlation circle.

- **Vectors**: Activate this option to display the initial variables in the form of vectors.
- **Orientate labels**: This option (only available with Excel 2010 and later versions) allows to orientate labels of variables in order to display them in the continuity of the vector.
- **Colored labels**: Activate this option to show labels in the same color as the points.
- **Color by group**: Activate this option, if you want to color variable points according to levels of a qualitative variable. Then select a vertical series of data that has as many rows as there are active variables. If headers were selected for the main table, a header must be included in this selection.
- **Resize points with Cos2**: Activate this option so that the variable points sizes are proportional to the respective sum of the squared cosines within the selected subspace.

Filter: Activate this option to modulate the number of variables displayed:

- **Random**: The observations to display are randomly selected. The "Number of variables" N to display must then be specified.
- **N first variables**: The first N variables are displayed on the chart. The "Number of variables" N to display must then be specified.
- **N last variables**: The last N variables are displayed on the chart. The "Number of variables" N to display must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the variables to display.
- **Sum(Cos2)>**: Only the variables for which the sum of squared cosines (communalities) are bigger than a value to enter are displayed on the plots.

Observations sub-tab:

Observations charts: Activate this option to display charts representing the observations in the new space.

- **Labels**: Activate this option to have observation labels displayed on the charts. The number of labels displayed can be changed using the filtering option.

- **Colored labels**: Activate this option to show labels in the same color as the points.
- **Color by group**: Activate this option, if you want to color observation points according to levels of a qualitative variable. Then select a vertical vector that must have as many rows as there are active observations. If headers were selected for the main table, ensure that a label is also present for the variable in this selection.
- **Confidence ellipses**: Activate this option if you want to display confidence ellipses around group of observations corresponding to the levels of the group variable selected to color observations. You also have to select the confidence interval for the ellipses.
- **Resize points with Cos2**: Activate this option so that the observation points sizes are proportional to the sum of the corresponding squared cosines within the selected subspace.

Filter: Activate this option to modulate the number of observations displayed:

- **Random**: The observations to display are randomly selected. The "Number of observations" N to display must then be specified.
- **N first rows**: The first N observations are displayed on the chart. The "Number of observations" N to display must then be specified.
- **N last rows**: The last N observations are displayed on the chart. The "Number of observations" N to display must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to display.
- **Sum(Cos2)>**: Only the observations for which the sum of squared cosines (communalities) are bigger than a value to enter will be displayed on the plots.

Biplots sub-tab:

Biplots: Activate this option to display charts representing the observations and variables simultaneously in the new space.

- **Options for variables**:
- **Vectors**: Activate this option to display the initial variables in the form of vectors.
- **Labels**: Activate this option to have variable labels displayed on the biplots.
- **Options for observations**:
- **Labels**: Activate this option to have observation labels displayed on the biplots.
- **Shared options**:

- **Supp. Obs/Var**: If you have included supplementary observations or supplementary variables in the PCA, activate this option to display them on the biplot.
- **Filter Obs/Var**: If you used a filter variable to display observations or variables, the same filter variable will be used to filter the display of observations and/or variables on the biplot.
- **Color Obs/Var**: If you used a group variable to color observations and/or variables, this same group variable will be used to color observations and/or variables on the biplot.

Type of biplot: Choose the type of biplot you want to display. See the description section for more details.

- **Correlation biplot**: Activate this option to display correlation biplots.
- **Distance biplot**: Activate this option to display distance biplots.
- **Symmetric biplot**: Activate this option to display symmetric biplots.
- **Coefficient**: Choose the coefficient whose square root is to be multiplied by the coordinates of the variables. This coefficient lets you to adjust the position of the variable points in the biplot in order to make it more readable. If set to other than 1, the length of the variable vectors can no longer be interpreted as standard deviation (correlation biplot) or contribution (distance biplot).

Bootstrap charts sub- tab:

Bootstrap observations chart: Activate this option to display charts containing the observations generated with the partial bootstrap method. See the description section for more details.

- **Number of samples**: Enter here the number of bootstrap samples to generate.
- **Color observations**: Activate this option so that each observation is colored with a different color.
- **Filter observations**: If you used a filter variable to display observations, the same filter variable will be used to filter the display of observations on the bootstrap chart.
- **Convex hulls**: Activate this option to display the convex hull corresponding to the bootstrap generated points.
- **Confidence ellipses**: Activate this option to display the confidence ellipse corresponding to the bootstrap generated points.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. This includes the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

Correlation/Covariance matrix: This table shows the data to be used afterwards in the calculations. The type of correlation depends on the option chosen in the "General" tab in the dialog box. For correlations, significant correlations are displayed in bold.

Bartlett's sphericity test: The results of the Bartlett sphericity test are displayed. They are used to confirm or reject the hypothesis according to which the variables do not have significant correlation.

Measure of Sample Adequacy of Kaiser-Meyer-Olkin: This table gives the value of the KMO measure for each individual variable and the overall KMO measure. The KMO measure ranges between 0 and 1. A low value corresponds to the case where it is not possible to extract synthetic factors (or latent variables). In other words, observations do not bring out the model that one could imagine (the sample is "inadequate"). Kaiser (1974) recommends not to accept a factor model if the KMO is less than 0.5. If the KMO is between 0.5 and 0.7 then the quality of the sample is mediocre, it is good for a KMO between 0.7 and 0.8, very good between 0.8 and 0.9 and excellent beyond.

Eigenvalues: The eigenvalues and corresponding chart (*scree plot*) are displayed. The number of eigenvalues is equal to the number of non-null eigenvalues.

If the corresponding output options have been activated, XLSTAT afterwards displays the **factor loadings** in the new space, then the correlations between the initial variables and the components in the new space. The **correlations** are equal to the factor loadings in a normalized PCA (on the correlation matrix).

If supplementary variables have been selected, the corresponding coordinates and correlations are displayed at the end of the table.

Contributions: Contributions are an interpretation aid. The variables which had the highest influence in building the axes are those whose contributions are highest.

Axes homogeneity index: This index developed by our team is very useful to determine if the contributions of the observations are homogeneous for the different axes. It is constructed as the proportion of observations with an absolute contribution > 1/n. An index above 0.4 indicates a very good homogeneity with well represented observations. On the other hand, an index lower than 0.1 should be a warning to the user who should check if there are no outliers in the variables constructing the axis that would distort its interpretation (the outliers would then be the observations that stand out from the others on the axis in question).

Squared cosines: As in other factor methods, squared cosine analysis is used to avoid interpretation errors due to projection effects. If the squared cosines associated with the axes used on a chart are low, the position of the observation or the variable in question should not be interpreted.

The **factor scores** in the new space are then displayed. If supplementary data have been selected, these are displayed at the end of the table. At the end of the factor scores table, the following button is displayed:  . Click on this button to automatically open the pre-filled dialog box of HAC (Hierarchical Ascending Classification) and perform a classification of the

observations on the factor scores.

Contributions: This table shows the contributions of the observations in building the principal components.

Squared cosines: This table displays the squared cosines between the observation vectors and the factor axes.

Where a rotation has been requested, the results of the rotation are displayed with the **rotation matrix** first applied to the factor loadings. This is followed by the modified variability percentages associated with each of the axes involved in the rotation. The coordinates, contributions and cosines of the variables and observations after rotation are displayed in the following tables.

Example

A tutorial on how to use Principal Component Analysis is available on the XLSTAT Help Center:

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

A tutorial on how to use Principal Component Analysis and apply filters based on communalities (squared cosines) is available on the XLSTAT Help Center:

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

References

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. Monographs on Statistics and Applied Probability, **54**, 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.

Morineau A. and Aluja-Banet T. (1998). Analyse en Composantes Principales. CISIA-CERESTA, Paris.

Factorial analysis of mixed data (PCAmix)

Use Factorial analysis of mixed data (PCAmix) to analyze a data table where observations are described both by quantitative variables and qualitative variables. This method is used to :

- Study and visualize correlations between variables.
- Obtain non-correlated factors which are linear combinations of the initial variables.
- Visualize observations in a 2- or 3-dimensional space.

In this section:

Description

The factorial analysis of mixed data is a method initially developed by Hill and Smith (1972). Few variants of this method were then developed (Escofier 1979, Pagès 2004). The method used in XLSTAT is called PCAmix and has been developed by Chavent et al (2014). This method can be seen as a mixture of two well known methods of factorial analysis: principal component analysis (PCA) which allows to study an observations/quantitative variables table and the multiple correspondence analysis (MCA) which allows to study an observations/qualitative variables table. The PCAmix method can be seen as a mixture of these two methods, it allows the analysis of a table where n observations are described by p_1 quantitative variables and by p_2 qualitative variables with a total of m_Q categories. We note $p = p_1 + p_2$. As the other factorial analysis methods, the PCAmix method aims to reduce data dimensionality as well as to identify nearness between variables but also proximity between the observations.

PCAmix results

The PCAmix method supplies the same classic results as other factorial analysis methods: factorial coordinates, contributions and squared cosines. These results are interpreted in the same way as in PCA or MCA. **Contributions** are a help to the interpretation. Variables having the most influenced the construction of axes are the ones whose contributions are the highest. **Squared cosines** allow to measure the quality of projection on a factorial axis and so to avoid errors of interpretation due to projection effects. If squared cosines associated to axes used on a graph are low, we shall avoid interpreting the position of the observation or the variable in question.

A specific output of PCAmix is called **squared loadings** of variables. The squared loading between a quantitative variable and a factorial axis is equal to the squared correlation between

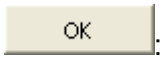
the variable and the axis k. The squared loading between a qualitative variable y and a factorial axis k is equal to the correlation ratio η^2 between the variable y and the axis k. We have:

$$\eta^2(k|y) = \frac{\sum_{s=1}^{m} n_s(\bar{k_s} - \bar{k})^2}{\sum_{i=1}^{n}(k_i - \bar{k})^2}$$

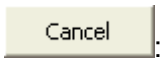
Where m is the total number of categories of the variable y, n_s is the number of observations having the categorie s, $\bar{k_s}$ is the mean of the variable k calculated on the observations having the categorie s and $\bar{k}$ is the mean of the variable k calculated on all the observations. This correlation ratio is equal to the sum of the contributions to the axis of each s categories of the qualitative variable y. Squared loadings are used to visualize quantitative variables and qualitative variables on the same chart and so their links with factorial axes.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Observations/quantitative variables: Select a table where n observations are described by p_1 quantitative variables. If column headers have been selected, check that the "Variable labels" option has been activated.

Observations/qualitative variables: Select a table where n observations are described by p_2 qualitative variables. If column headers have been selected, check that the "Variable labels" option has been activated.

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (observations/variables table, weights, observation labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Options tab:

Sort categories alphabetically: Activate this option so that the categories of all the variables are sorted alphabetically.

Variable-Category labels: Activate this option to use variable-category labels when displaying outputs. Variable-Category labels include the variable name as a prefix and the category name as a suffix.

Filter factors: You can activate one of the three following options in order to reduce the number of factors displayed:

- **Minimum %**: Activate this option and then enter the minimum percentage that should be reached to determine the number of factors to display.
- **Maximum number**: Activate this option to set the maximum number of factors to take into account when displaying the results.

Supplementary data tab:

Supplementary observations: Activate this option if you want to represent additional observations by calculating their coordinates. These observations are neither taken into account for the computation of the correlation matrix, nor for the subsequent calculations (we talk of passive observations as opposed to active observations). If the first row of the data selection for supplementary observations includes a header you must activate the "Variable labels for supp. obs" option. You can also select labels for supplementary observations which will be used for the display.

- **Quantitative variables**: select a table with n' supplementary observations described by the same p_1 active quantitative variables.
- **Qualitative variables**: select a table with n' supplementary observations described by the same p_2 active qualitative variables.

Supplementary variables: Activate this option if you want to compute a posteriori the coordinates of variables that are not taken into account for the computing of the principal axes (passive variables, as opposed to active variables).

- **Quantitative**: Activate this option if you want to include quantitative supplementary variables. If the headers of the columns of the main table have been selected, you also need to select headers here.

- **Qualitative**: Activate this option if want to include qualitative supplementary variables. If the headers of the columns of the main table have been selected, you also need to select headers here.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to ignore the observations that contain missing data.

Replace missing data: Activate this option to replace missing data. When a missing data corresponds to a quantitative variable, they are replaced by the mean of the variable. When a missing data corresponds to a qualitative variable , a new "Missing" category is created for the variable.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics for the selected variables.

Eigenvalues: Activate this option to display the table and the scree plot of the eigenvalues.

Display results for:

- **Observations and variables**: Activate this option to display the results that concern the observations and the variables.

- **Observations**: Activate this option to display only the results that concern the observations.

- **Variables**: Activate this option to display only the results that concern the variables.

Principal coordinates: Activate this option to display the principal coordinates.

Contributions: Activate this option to display the contributions.

Squared cosines: Activate this option to display the squared cosines.

Squared loadings: Activate this option to display the squared loadings.

Charts tab:

Quantitative sub-tab:

Correlations charts: Activate this option to display charts showing the correlations between the components and initial variables. This chart is named correlation circle.

- **Vectors**: Activate this option to display the initial variables in the form of vectors.
- **Orientate labels**: This option (only available with Excel 2010 and later versions) allows to orientate labels of variables in order to display them in the continuity of the vector.
- **Colored labels**: Activate this option to show labels in the same color as the points.
- **Color by group**: Activate this option, if you want to color variable points according to levels of a qualitative variable. Then select a vertical series of data that has as many rows as there are active variables. If headers were selected for the main table, a header must be included in this selection.
- **Resize points with Cos2**: Activate this option so that the variable points sizes are proportional to the respective sum of the squared cosines within the selected subspace.

Filter: Activate this option to modulate the number of variables displayed:

- **Random**: The observations to display are randomly selected. The "Number of variables" N to display must then be specified.
- **N first variables**: The first N variables are displayed on the chart. The "Number of variables" N to display must then be specified.
- **N last variables**: The last N variables are displayed on the chart. The "Number of variables" N to display must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the variables to display.
- **Sum(Cos2)>**: Only the variables for which the sum of squared cosines (communalities) are bigger than a value to enter are displayed on the plots.

Qualitative sub-tab:

Factorial map of categories: Activate this option to display the chart showing the principal coordinates of categories of active and supplementary qualitative variables.

- **Labels**: Activate this option to show labels of categories on the chart.
- **Colored labels**: Activate this option to show labels in the same color as the points.
- **Color by group**: Activate this option, if you want to color variable points according to levels of a qualitative variable. Then select a vertical series of data that has as many rows as there are active variables. If headers were selected for the main table, a header must be included in this selection.
- **Resize points with Cos2**: Activate this option so that the categories points sizes are proportional to the respective sum of the squared cosines within the selected subspace.
- **Link categories**: Activate this option so that the categories belonging to a given variable are linked. This option allows to quickly distinguish categories which belong to a variable.

Filter: Activate this option to modulate the number of variables displayed:

- **Random**: The observations to display are randomly selected. The "Number of variables" N to display must then be specified.
- **N first variables**: The first N variables are displayed on the chart. The "Number of variables" N to display must then be specified.
- **N last variables**: The last N variables are displayed on the chart. The "Number of variables" N to display must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the variables to display.

Mixed sub-tab:

Mixed chart: Activate this option to display the mixed chart of squared loadings for quantitative and qualitative variables.

Options for quantitative variables:

- **Labels**: Activate this option to have quantitative variables labels displayed on the mixed chart.
- **Vectors**: Activate this option to display the quantitative variables in the form of vectors.

- **Supplementary variables**: If you have included supplementary quantitative variables, activate this option to display them on the mixed chart.
- **Filer variables**: If you used a filter variable to display quantitative variables, the same filter variable will be used to filter the display of quantitative variables on the mixed chart.

Options for qualitative variables:

- **Labels**: Activate this option to have qualitative variables labels displayed on the mixed chart.
- **Vectors**: Activate this option to display the qualitative variables in the form of vectors.
- **Supplementary variables**: If you have included supplementary qualitative variables, activate this option to display them on the mixed chart.
- **Filer variables**: If you used a filter variable to display qualitative variables, the same filter variable will be used to filter the display of quantitative variables on the mixed chart.

Observations sub-tab:

Factorial map of observations: Activate this option to display charts representing the principal coordinates of observations.

- **Labels**: Activate this option to have observation labels displayed on the charts. The number of labels displayed can be changed using the filtering option.
- **Colored labels**: Activate this option to show labels in the same color as the points.
- **Color by group**: Activate this option, if you want to color observation points according to levels of a qualitative variable. Then select a vertical series of that that must have as many rows as there are active observations. If headers were selected for the main table, ensure that a label is also present for the variable in this selection.
- **Resize points with Cos2**: Activate this option so that the observation points sizes are proportional to the sum of the corresponding squared cosines within the selected subspace.

Colors tab:

This tab allows you to personalize colors of differents numerical results and charts colors.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected

Eigenvalues: The eigenvalues and corresponding chart (*scree plot*) are displayed. The number of eigenvalues is equal to the number of non-null eigenvalues.

Variables results: If the options of corresponding outputs have been activated, XLSTAT displays outputs for active variables (principal coordinates, squared cosines, contributions and squared loadings). The first part of every table concerns quantitative variables, the second part concerns qualitative variables. If supplementary variables have been selected, results for these variables are then shown. The correlation circle of quantitative variables, the factorial map of the categories of the qualitative variables and the mixed chart of squared loadings are then shown.

Observations results: If the options of corresponding outputs have been activated, XLSTAT displays outputs for active observations (principal coordinates, squared cosines and contributions). If supplementary observations have been selected, results for these observations are then shown. The factorial map of the observations is then shown.

At the end of the observations coordinates table (factor scores), the following button is displayed:  . Click on this button to automatically open the pre-filled dialog box of HAC (Hierarchical Ascending Classification) and perform a classification of the observations on the factorial coordinates.

*Remark about the **Axes homogeneity index***: This index developed by our team is very useful to determine if the contributions of the observations are homogeneous for the different axes. It is constructed as the proportion of observations with an absolute contribution > 1/n. An index above 0.4 indicates a very good homogeneity with well represented observations. On the other hand, an index lower than 0.1 should be a warning to the user who should check if there are no outliers in the variables constructing the axis that would distort its interpretation (the outliers would then be the observations that stand out from the others on the axis in question).

Example

A tutorial on how to use PCAmix is available on the XLSTAT Help Center:

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

References

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.

Discriminant Analysis (DA)

Use discriminant analysis to explain and predict the membership of observations to several classes using quantitative or qualitative explanatory variables.

In this section:

Description

Discriminant Analysis (DA) is an old method (Fisher, 1936) which in its classic form has changed little in the past twenty years. This method, which is both explanatory and predictive, can be used to:

- Check on a two- or three-dimensional chart if the groups to which observations belong are distinct,
- Show the properties of the groups using explanatory variables,
- Predict which group an observation will belong to.

DA may be used in numerous applications, for example in ecology and the prediction of financial risks (credit scoring).

The principle of DA is to model a qualitative dependent variable Y from a new variable called discriminant. The latter one is a linear combination of the explanatory variables and is chosen so that it best discriminates between the classes defined by the modalities of the variable to be explained.

Let be n individual described by p explanatory variables x_i = (x_i^1, ..., x_i^n), the goal is to find a discriminant variable $s \in R^n$ linear combination of vectors x_1, ..., x_p such that: $$ s = u_1 x_1 + . . . + u_p x_p , $$

where u = (u_1, . . . , u_p) $\in R^p$ is the vector of the coefficients of this linear combination, called the discriminating factor. For that we must find on one hand the vector u and on the other hand a metric to justify that s discriminates between classes.

Consider the total variance-covariance matrix T whose elements are written: $$ t ^ { j, l } = \frac{ 1 }{ n } \sum _ {i = 1}^{i = n} (x _ { i }^{ j } - \mu ^ { j })(x _ { i }^{ l } - \mu ^{ l }) , j, l = 1, ..., p, \tag{1} $$

assuming that all individuals are the same weight and where $$ \mu ^ {j} = \frac{ 1 }{ n } \sum_{i = 1}^{i = n} x_{ i }^{ j }, j= 1,..., p, $$

is the average of the observations on the variable x_j. AFD is based on the analysis of variance, its decomposition is made on a partition of the data as follows: $$ t ^ { j, l } = \frac{ 1 }{ n } \sum_{k = 1}^{k = g}\sum_{x_{ i } \in G_{ k }} (x _ { i }^{ j } - \mu_{ k } ^{ j } + \mu_{ k }^{ j } - \mu^{ j })(x _ { i }^{ l } - \mu_{ k }^{ l } + \mu_{ k }^{ l } - \mu ^{ l }) , j, l = 1, ..., p.\tag{2} $$

where μ_k is the average of the class G_k, for k ranging from 1 to g, the number of classes given by the dependent variable. By developing the product of equation (2) (see Bardos (2001) for more details), we deduce that the matrix T breaks down into the sum of 2 matrices: B, the inter-class variance-covariance matrix of g centers of gravity μ_k, and W, the intra-class variance-covariance matrix resulting from the sum of the matrices W_k obtained for each class G_k.

A good class discrimination is possible if the centers of gravity projected into the space R^n are far apart and if the projected groups are not too dispersed. This is possible by maximizing the ratio between the inter-class variance and the total variance, i.e. we seek u, such that $$ \max_{u \in R^{ p }} \frac{u^{ ' }Bu}{u^{ ' }Tu}. $$

By definition this ratio is maximal if u is an eigenvector of $T^{-1}B$ associated with the greatest eigenvalue. The number of non-zero eigenvalues is at most equal to $(g-1)$ where g is the number of classes.

The score function then allows you to calculate for each individual x his membership in a class. For that we come back to the calculation of the distances compared to the centers of the classes and to their comparison. The function is written with the notations above: $$ f_ { i, k } (x) = \frac{1}{2} \delta_{i / k}(x) = (\mu _ { i } - \mu _ { k })^{ ' } M \left(x - \frac{\mu _ { i } + \mu _ { k }}{2}\right) , i, k = 1, ..., g\tag{3} $$

where M is a positive definite symmetric matrix and the $\left\{\left(\mu_i-\mu_k\right)^{'} M\right\}_{i,k}$ represent the coefficients of the canonical discriminant function. We then have g distances, ie one per group. For an individual x the set of distances is calculated and compared. The individual x will be assigned to the group that gives the smallest distance.

SSCP matrices and distance matrices

The XLSTAT DA feature proposes to calculate the SSCP matrices or matrix of sums of squares and cross products. They are constructed like the covariance matrices and are proportional to them. They also verify the following relation: total SSCP = inter SSCP + total intra SSCP. In particular, the intra SSCP matrix is used in calculations such as for example in statistical tests or the calculation of the coefficients of the canonical discriminant function ($M = W^{-1}$).

When the matrix M of equation (3) is replaced by the inverse of the intra-class variance-covariance matrix W^{-1} then we fall back on the Mahalanobis distance between x and μ_i. A detailed description of these matrices is given in the help [here] (MAH.md). In the case where the intra-class variance-covariance matrices are assumed to be equal, the distance matrix is calculated using the total intra-class covariance matrix.

In the case of the assumption of equality of the covariance matrices, the Fisher distances between the classes are calculated. They are obtained from the Mahalanobis distance and allow a test of significance. If we do not assume that the covariance matrices are equal, the generalized quadratic distances between the classes are proposed in the results. The generalized distance is also calculated from the Mahalanobis distances and takes into account the logarithms of the determinants of the covariance matrices as well as the logarithms of the prior probabilities.

Linear or quadratic model

Two models of DA are used depending on a fundamental assumption: If the covariance matrices are assumed to be identical, linear discriminant analysis is used. If, on the contrary, it is assumed that the covariance matrices differ between at least two classes, then the quadratic model is used. This fundamental assumption is proposed in the options of the DA dialog box. If you do not know which option to choose, the Box test is displayed in the output options to test this hypothesis. Bartlett's approximation which is based on a law of χ^2 also makes it possible to perform this test. To better evaluate this test, it is suggested to start by performing a linear analysis, then, depending on the results of Box's test, possibly perform a quadratic analysis.

Multicolinearity issues

With linear and, even more, with quadratic models, we can face problems of variables with a null variance or multicolinearity issues between variables. XLSTAT has been programmed so as to deal with these problems. The variables responsible for these problems are automatically ignored either for all calculations or, in the case of a quadratic model, for the groups in which the problems arise. Multicolinearity statistics are optionally displayed so that you can identify the variables which are causing problems.

Stepwise methods

As for linear regression and logistic regression, efficient step-by-step methods have been proposed. However, they can only be used when only quantitative variables are selected because the input and output tests of variables are based on the normality assumption of the variables. The stepwise method allows to obtain an efficient model avoiding variables which provide little information to the model. These methods are available in the options tab of the DA dialog box.

Classification table, ROC curve and cross-validation

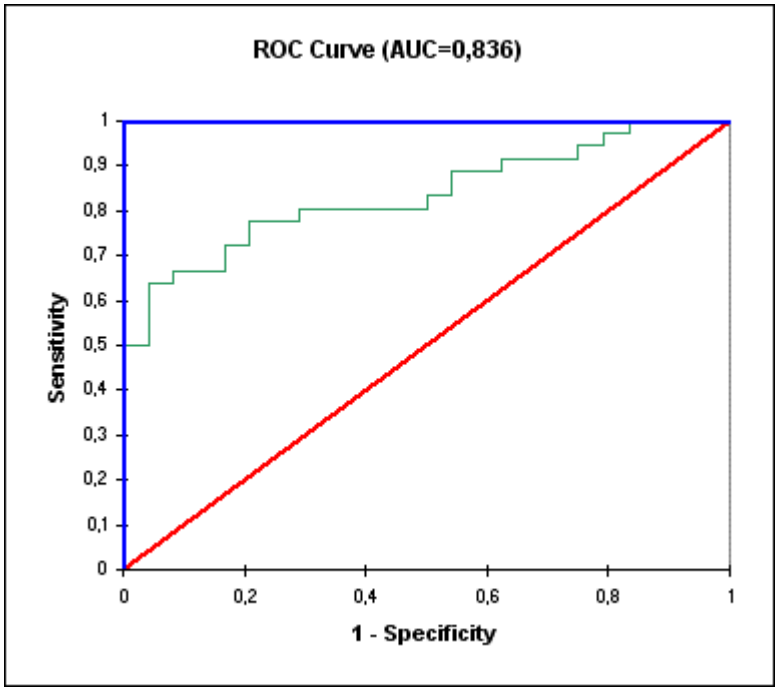
Among the numerous results provided, XLSTAT can display the classification table (also called confusion matrix) used to calculate the percentage of well-classified observations. When only two classes (also called categories or modalities) are present in the dependent variable, the ROC curve can be also displayed.

The ROC curve (*Receiver Operating Characteristics*) displays the performance of a model and enables a comparison to be made with other models. The terms used come from signal detection theory.

The proportion of well-classified positive events is called the sensitivity. The specificity is the proportion of well-classified negative events. If you vary the threshold probability from which an

event is to be considered positive, the sensitivity and specificity will also vary. The curve of points (1-specificity, sensitivity) is the ROC curve.

Let's consider a binary dependent variable which indicates, for example, if a customer has responded favorably to a mail shot. In the diagram below, the blue curve corresponds to an ideal case where the $n\%$ of people responding favorably corresponds to the $n\%$ highest probabilities. The green curve corresponds to a well-discriminating model. The red curve (first bisector) corresponds to what is obtained with a random Bernoulli model with a response probability equal to that observed in the sample studied. A model close to the red curve is therefore inefficient since it is no better than random generation. A model below this curve would be disastrous since it would be less even than random.



The area under the curve (or *AUC*) is a synthetic index calculated for ROC curves. The AUC corresponds to the probability such that a positive event has a higher probability than a negative event. For an ideal model, $AUC = 1$ and for a random model, $AUC = 0.5$. A model is usually considered good when the $AUC > 0.7$. A well-discriminating model must have an AUC of between 0.87 and 0.9. A model with an $AUC > 0.9$ is excellent.

The results of the model as regards forecasting may be too optimistic: we are effectively trying to check if an observation is well-classified while the observation itself is being used in calculating the model. For this reason, cross-validation was developed: to determine the probability that an observation will belong to the various groups, it is removed from the learning sample, then the model and the forecast are calculated. This operation is repeated for all the observations in the learning sample. The results thus obtained will be more representative of the quality of the model. XLSTAT gives the option of calculating the various statistics associated with each of the observations in cross-validation mode together with the classification table and the ROC curve if there are only two classes.

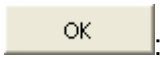
Lastly, you are advised to validate the model on a validation sample wherever possible. XLSTAT has several options for generating a validation sample automatically.

Discriminant analysis and logistic regression

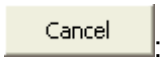
In the case where there are only two classes to predict for the variable dependent, the discriminant analysis is very close to the regression logistics. The discriminant analysis has the advantage of studying in the detail the covariance structures, and lead to a representation graphic. Logistic regression has the advantage of offering several forms of possible models, and the use of step-by-step selection methods for the qualitative explanatory variables. The user will be able to compare the performances of the two methods by relying on the ROC curves.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: click this button to start the calculations.



: click this button to close the dialog box without doing any calculations.

: click this button to display help.


: click this button to reload the default options.


: click this button to delete the data selections.


: click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.


: click this button to change the way you load data into XLSTAT. If the button has a mouse icon, XLSTAT allows you to make a mouse selection of the data. If the button has a list icon, XLSTAT allows you to select data from a list. If the button has an orange icon, additional buttons with a question mark! [Select_file_choosefile.png] (img/Select_file_choosefile.png) {: width = "26" height = "25"} are displayed to allow you to import data from files.

General tab:

Y / Dependent variables:

Qualitative: Select the qualitative variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

X / Explanatory variables:

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected must be of the numerical type. If a variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data must be of any type, but numerical data will automatically be considered as nominal. If a variable header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observation labels) includes a header.

Observation labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observation labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be considered as 1. XLSTAT uses these weights for calculating degrees of freedom. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option has been activated.

Options tab:

Tolerance: Enter the value of the tolerance threshold below which a variable will automatically be ignored.

Equality of covariance matrices: Activate this option if you want to assume that the covariance matrices associated with the various classes of the dependent variable are equal.

Prior probabilities: Activate this option if you want to take prior possibilities into account. The probabilities associated with each of the classes are equal to the frequency of the classes. Note: this option has no effect if the prior possibilities are equal for the various groups.

Filter factors: You can activate one of the two following options in order to reduce the number of factors used in the model:

- **Minimum %**: Activate this option and enter the minimum percentage of total variability that the selected factors should represent.
- **Maximum number**: Activate this option to set the maximum number of factors to take into account.

Significance level (%): Enter the significance level for the various tests calculated.

Model selection: Activate this option if you want to use one of the four selection methods provided:

- **Stepwise (Forward)**: The selection process starts by adding the variable with the largest contribution to the model. If a second variable is such that its entry probability is greater than the **entry threshold value**, then it is added to the model. After the third variable is added, the impact of removing each variable present in the model after it has been added is evaluated. If the probability of the calculated statistic is greater than the **removal threshold value**, the variable is removed from the model.
- **Stepwise (Backward)**: This method is similar to the previous one but starts from a complete model.
- **Forward**: The procedure is the same as for stepwise selection except that variables are only added and never removed.
- **Backward**: The procedure starts by simultaneously adding all variables. The variables are then removed from the model following the procedure used for stepwise selection.

Classes weight correction: If the number of observations for the various classes for the dependent variables are not uniform, there is a risk of penalizing classes with a low number of observations in establishing the model. To get over this problem, XLSTAT has two options:

- **Automatic**: Correction is automatic. Artificial weights are assigned to the observations in order to obtain classes with an identical sum of weights.
- **Corrective weights**: You can select the weights to be assigned to each observation.

Validation tab:

Validation: Activate this option if you want to use a sub-sample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.

- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: The first row of the selections listed below must correspond to data.

Quantitative: Activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

Qualitative: Activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

Observation labels: Activate this option if observation labels are available. Then select the corresponding data. If this option is not activated, the observation labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

General sub-tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the correlation matrix.

Means by class: Activate this option to display the means of each class for each of the explanatory variables.

Information on each class: Activate this option to display information relating to each class, namely the mean, the sum of the weights, the probability of assignment and the logarithm of the determinant calculated on the covariance matrix.

Multicolinearity statistics: Activate this option to display the table of multicolinearity statistics.

SSCP matrices: Activate this option to display the sums of squares and cross-product matrices for each explanatory variable (intra-class) and, optionally, each interaction (inter-class) and total.

Covariance matrices: Activate this option to display inter-class, intra-class, total intra-class, and total covariance matrices.

Distance matrices: Activate this option to display the matrix of distances between classes.

Eigenvalues: Activate this option to display the table of eigenvalues.

Eigenvectors: Activate this option to display the eigenvector table.

Variables/Factors correlations: Activate this option to display correlations between factors and variables.

Canonical correlations and functions: Activate this option for canonical correlations and functions.

Classification functions: Activate this option to display classification functions.

Scores: Activate this option to display the coordinates of the observations in the factor space. The prior and posterior classes for each observation, the probabilities of assignment for each class and the distances of the observations from their centroid are also displayed in this table.

Confusion matrix: Activate this option to display the table showing the numbers of well- and badly-classified observations for each of the classes.

Cross-validation: Activate this option to display cross-validation results (probabilities for observations and confusion matrix).

Tests sub-tab:

Means by class: Activate this option to display the tests related on the means by class.

- **Wilks lambda test (Rao approximation)**: Activate this option to display the results of the Lambda statistic and the associated p-value.
- **Pillai's trace**: Activate this option to display the results of the Pillai trace test.
- **Hotelling-Lawley trace**: Activate this option to display the results of the Hotelling-Lawley trace test.
- **Largest Roy Root**: Activate this option to display the results of the largest Roy root test.

Within-class covariance matrices: Activate this option to display the tests on the intra-class covariance matrices.

- **Box test**: Activate this option to display the results of the Box test and the p-values resulting from the two possible approximations.
- **Kullback test**: Activate this option to display the results of the Kullback test.

Eigenvalues: Activate this option to display the test related on the eigenvalues.

- **Bartlett test**: Activate this option to display the results of the Bartlett test results. This test is possible only if the number of eigenvalues is strictly greater than 1.

Charts tab:

Correlation charts: Activate this option to display the charts involving correlations between the factors and input variables.

Eigenvalues: Activate this option to display the chart of the Eigenvalues (scree plot).

Observations charts: Activate this option to display the charts that allow visualizing the observations in the Eigenvectors space.

- **Labels**: Activate this option to display the observation labels on the charts. The number of labels can be modulated using the filtering option.
- **Display the centroids**: Activate this option to display the centroids that correspond to the categories of the dependent variable.
- **Confidence ellipses**: Activate this option to display confidence ellipses. The confidence ellipses correspond to a $x\%$ confidence interval (where x is determined using the significance level entered in the Options tab) for a bivariate normal distribution with the same means and the same covariance matrix as the factor scores for each category of the dependent variable.

Centroids and confidence circles: Activate this option to display a chart with the centroids and the confidence circles around the means.

Colored labels: Activate this option to display the labels with the same color as the corresponding points. If this option is not activated the labels are displayed in black.

Filter: Activate this option to modulate the number of observations displayed:

- **Random**: The observations to display are randomly selected. The "Number of observations" N to display must then be specified.
- **N first rows**: The N first observations are displayed on the chart. The "Number of observations" N to display must then be specified.

- **N last rows**: The N last observations are displayed on the chart. The "Number of observations" N to display must then be specified.

- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to display.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. This includes the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

Correlation matrix: This table displays the correlations between the explanatory variables.

Means by class: This table provides the means of the various explanatory variables for the various classes of the dependent variable.

Unidimensional test of equality of the means of the classes: This table provides the results of the unidimensional test of equality of class means which tests, variable by variable, the hypothesis of equality of means between classes. Wilks' univariate lambda is always between 0 and 1. A value of 1 means the class means are equal. A low value is interpreted as meaning there are low intra-class variations and therefore high inter-class variations, hence a significant difference in class means.

Wilks' Lambda test (Rao's approximation): This table provides the results of Wilks' Lambda test which tests the hypothesis of equality of the mean vectors for the various classes. When there are two classes, the test is equivalent to the Fisher test mentioned previously. If the number of classes is less than or equal to three, the test is exact. The Rao approximation is required from four classes to obtain a statistic approximately distributed according to a Fisher distribution. A description of this test is given in the help [here] (MNV.md).

Pillai's trace: This table provides the results of Pillai's trace test which tests the hypothesis of equality of the mean vectors for the various classes. It is less used than Wilks' Lambda test and also uses the Fisher distribution for calculating p-values. A description of this test is given in the help [here] (MNV.md).

Hotelling-Lawley trace: This table provides the results of Hotelling-Lawley trace test which tests the hypothesis of equality of the mean vectors for the various classes. It is less used than Wilks' Lambda test and also uses the Fisher distribution for calculating p-values. A description of this test is given in the help [here] (MNV.md).

Roy's greatest root: This table provides the results of Roy's greatest root test which tests the hypothesis of equality of the mean vectors for the various classes. It is less used than Wilks' Lambda test and also uses the Fisher distribution for calculating p-values. A description of this test is given in the help [here] (MNV.md).

Sum of weights, prior probabilities and logarithms of determinants for each class: This table provides the sum of the weights, the probability of membership and the logarithm of the determinant for each covariance matrix. These statistics are used, among other things, in the calculations of posterior probabilities for observations.

Multicolinearity: This table provides the list of the variables responsible for multicollinearities between the variables. As soon as a variable is detected as being responsible for multicollinearity (its tolerance is less than the limit tolerance set in the "options" tab of the dialog box), it is not included in the multicolinearity statistics calculation for the following variables. Thus in an extreme case where two variables are identical, only one of the two variables will be eliminated from the calculations. The statistics displayed are the tolerance (equal to $1 - R^2$), and its inverse, the VIF (Variance inflation factor).

SSCP matrices: This result provides the SSCP matrices (*Sums of Squares and Cross Products*) inter, intra and total successively.

Covariance matrices: This result provides the inter-class variance-covariance matrix, the intra-class variance-covariance matrices of each of the classes and the total one successively.

Summary of the variables selections: This table provides a summary of the variables selection in the case where a selection method has been chosen in the options. In the case of a stepwise method, ascending or descending selection, the statistics corresponding to the different steps are displayed.

Box test: This table provides the results of the Box test. Two approximations are available, one based on the χ^2 distribution, and the other on the Fisher distribution. The results of both tests are displayed.

Kullback's test: This table provides the results of the Kullback's test. The statistic calculated is approximately distributed according to a χ^2 distribution.

Mahalanobis distances: This table provides the Mahalanobis distances.

Fisher's distances: This table provides the Fisher's distances. The matrix of p-values is displayed so as to identify which distances are significant.

Generalized squared distances: This table provides the generalized squared distances between the classes.

Eigenvalues: This table provides the eigenvalues associated with the various factors together with the corresponding discrimination percentages and cumulative percentages. The sum of the eigenvalues is equal to the Hotelling trace. The scree plot is displayed below if the option has been chosen in order to visualize how the discriminating power is distributed among the discriminating factors.

Bartlett's test on significancy of eigenvalues: This table provides for each eigenvalue, the Bartlett statistic and the corresponding p-value which is computed using the asymptotic Chi-square approximation. The Bartlett's test allows to test the null hypothesis H_0 that all the p eigenvalues are equal to zero. If it is rejected for the greatest eigenvalue then the test is performed again until H_0 cannot be rejected. This test is known as conservative, meaning that it tends to confirm H_0 in some cases where it should not. You can however use this test to check how many factorial axes you should consider (see Jobson, 1992).

Eigenvectors: This table provides the eigenvectors afterwards used in the canonical correlations, canonical function coefficients and observation coordinate calculations.

Variables/Factors correlations: This table provides the correlations between the observation coordinates in the space of the initial variables and in the space of the discriminating factors. The circle of correlations is displayed below if the option has been chosen to ease the interpretation of the observation representation in factor space.

Canonical correlations: This table provides the canonical correlations associated with each factor. Canonical correlations are also a measurement of the discriminant power of the factors. Their sum is equal to the Pilai's trace.

Canonical discriminant function coefficients: This table provides the canonical discriminant function coefficients which are used to calculate the coordinates of an observation in discriminant factor space from its coordinates in the initial variable space.

Standardized canonical discriminant function coefficients: This table provides the standardized coefficients of those given above. Their comparison thus gives a measure of the relative contribution of the initial variables to the discrimination for a given factor.

Classification functions: This table provides the classification functions which are used to determine to which class should be assigned an observation on the basis of the values taken for the different explanatory variables. An observation is assigned to the class with the highest classification function.

Prior and posterior classification, membership probabilities, scores and squared distances: This table provides for each observation its membership class defined by the dependent variable, the membership class as deduced by the membership probabilities, the probabilities of membership of each of the classes, the coordinates in discriminant factor space and the squared distances of the observations from the centroids of each of the classes.

Confusion matrix for the estimation sample: This table provides the confusion matrix, as well as the overall percentage of well-classified observations. If the dependent variable only includes two classes, the ROC curve is displayed (see the description section for more details).

Cross-validation: This table provides, if the cross-validation option has been chosen, the information for the observations and the confusion matrix (see the description section for more details).

Example

A tutorial on how to use Discriminant Analysis is available on the XLSTAT Help Center:

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

References

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.

Correspondence Analysis (CA)

Use this tool to visualize the links between the categories of two qualitative variables. The variables can be available as an observations/variables table, as a contingency table, or as a more general type of two-way table.

In this section:

Description

Correspondence Analysis (CA) is a powerful method that lets you study the association between two qualitative variables. The research of J.-P. Benzécri that started in the early sixties led to the emergence of the method. His disciples worked on several developments of the basic method. For example, M.J. Greenacre's book (1984) contributed to the popularity of the method throughout the world. The work of C. Lauro from the University of Naples led to a non-symmetrical variant of the method.

Measuring the association between two qualitative variables is a complex subject that first requires transforming the data: it is not possible to compute a correlation coefficient using the data directly, as one could do with quantitative variables.

The first transformation consists of recoding the two qualitative variables V_1 and V_2 as two disjunctive tables Z_1 and Z_2 or indicator (or dummy) variables. For each category of a variable there is a column in the respective disjunctive table. Each time the category *c* of variable V_1 occurs for an observation *i*, the value of $Z_1(i,c)$ is set to one (the same rule is applied to the V_2 variable). The other values of Z_1 and Z_2 are zero. The generalization of this idea to more than two variables is called Multiple Correspondence Analysis. When there are only two variables, it is sufficient to study the contingency table of the two variables, that is the table $Z_1^t Z_2$.

The contingency table has the following structure:

$V_1 \setminus V_2$	**Category 1**	**…**	**Category j**	**…**	**Category** m_2
Category 1	n_{11}		n_{1j}	…	n_{1m_2}
…	…	…	…	…	…
Category i	n_{i1}	…	n_{ij}	…	n_{im_2}
…	…	…	…	…	…
Category m_1	$n_{m_1 1}$	…	$n_{m_1 j}$	…	$n_{m_1 m_2}$

where n_{ij} is the frequency of observations that show both characteristic i for variable V_1, and characteristic j for variable V_2.

The Chi-square distance has been suggested to measure the distance between two categories. The sum of these distances for the whole table is equal to the Chi-square statistic which asymptotically follows a Chi-square distribution with $(m_1 - 1)(m_2 - 1)$ degrees of freedom. This statistic allows us to test the hypothesis of independence between the rows and columns of the contingency table.

Inertia is a measure inspired from physics that is often used in Correspondence Analysis. The inertia of a set of points is the weighted mean of the squared distances to the center of gravity. In the specific case of CA, the total inertia of the set of points (one point corresponds to one category) can be written as:

$$\phi^2 = \frac{\chi^2}{n}$$

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

and where n is the sum of the frequencies in the contingency table. We can see that the inertia is proportional to the Pearson Chi-square statistic computed on the contingency table.

The aim of CA is to represent as much of the inertia on the first principal axis as possible, a maximum of the residual inertia on the second principal axis and so on until all of the total inertia is represented in the space of the principal axes. One can show that the number of dimensions of the space is equal to $min(m_1, m_2) - 1$.

Non-Symmetrical Correspondence Analysis (NSCA) developed by Lauro and D'Ambra (1984) analyzes the association between the rows and columns of a contingency table while introducing the notion of dependency between the rows and the columns, which leads to an asymmetry in their treatment. The example the authors used in their first article on this subject corresponds to the analysis of a contingency table that contains the prescriptions of 6 drugs for 7 different diseases for 69 patients. Here there is an obvious dependency of the drugs on the disease. In order to take into account the dependency, Goodman and Kruskal's tau (1954) was suggested. The tau coefficient that corresponds to the case where the rows depend on the columns can be written as:

$$\tau_{b/RC} = \frac{\sum_{j=1}^{m_2} \frac{n_{.j}}{n} \sum_{i=1}^{m_1} (\frac{n_{ij}}{n_{.j}} - \frac{n_{i.}}{n})^2}{1 - \sum_{i=1}^{m_1} \frac{n_{i.}}{n}^2}$$

As with the total inertia, it is possible to compute a representation space for the categories, such that the proportion of the Goodman and Kruskal's tau represented on the chart is maximized on the first axes.

Greenacre (1984) defined a framework (the generalized singular value decomposition) that allows computing both CA and the related method of NSCA in a similar way.

An alternative approach using the **Hellinger distance** was proposed by Rao (1995). The Hellinger distance only depends on the profiles of the concerned pair and does not depend on the sample sizes on which the profiles are estimated. Therefore, the Hellinger distance approach might be a good alternative to the classical CA when average column profiles are not relevant (e.g. when columns represent populations of individuals classified according to row categories) or if some categories have low frequencies. Computation follows the unified approach described by Cuadras and Cuadras (2008). The inertia is generalized by the following formula:

$$\phi^2(\alpha, \beta) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} [(\frac{\frac{n_{ij}}{n}}{\frac{n_{i.}}{n} \frac{n_{.j}}{n}})^\alpha - 1]^2 \frac{n_{i.}}{n} (\frac{n_{.j}}{n})^{2\beta}$$

Notes:

- In the case of Correspondence Analysis using the Hellinger distance, $\alpha = \frac{1}{2}$ and $\beta = \frac{1}{2}$.
- In the of classical Correspondence Analysis, $\alpha = 1$ and $\beta = \frac{1}{2}$

The **Analysis of a subset of categories** is a new method that was recently developed by Greenacre (2006). It allows parts of tables to be analyzed while maintaining the margins of the whole table and thus the same weights and Chi-square distances of the whole table, simplifying the analysis of large tables by breaking down the interpretation into parts.

The **Detrended Correspondence Analysis** (DCA) is a method proposed by Hill and Gauch (1980), mainly used on ecological data. The aim of this method is to correct two drawbacks encountered when using classical CA:

```
* The first is the "arc effect", also called the "horseshoe effect". This effec

* The second disadvantage often encountered is the tendency of CA to compress t
```

Note: DCA can only calculate coordinates on a maximum of 4 dimensions.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display help options.

: Click this button to reload the default options.

: Click this button to delete the data selections.

: Click these buttons to change the way XLSTAT loads the data:

- Case where the data are in a contingency table or a more general two-way table: if the arrow points down, XLSTAT allows you to select data by columns or by range. If the arrow points to the right, XLSTAT allows you to select data by rows or by range.
- Case where the data are in an observations/variables table: if the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 ![Select_list.png][image-9]{: width="26" height="25"} ![Select_file.png][image-10]{: width="26" height="25"} : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files ![Select_file_choosefile.png][image-11]{: width="26" height="25"}.

General tab:

The first selection field lets you alternatively select two types of tables:

Two-way table: Select this option if your data correspond to a two-way table where the cells contain the frequencies corresponding to the various categories of two qualitative variables (in this case it is more precisely a contingency table), or other type of values.

Observations/variables table: Select this option if your data correspond to N observations described by 2 qualitative variables. This type of table typically corresponds to a survey with 2 questions. During the computations, XLSTAT will automatically transform this table into a contingency table.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: This option is visible if the selected table is a contingency table or a more general two-way table. Activate this option if the labels of the columns and rows are included in the selection. In this case the first column of the selection contains the row labels and the first row contains the column labels.

Variable labels: This option is visible only if you selected the observations/variables table format. Activate this option if the first row contains the variable labels (case of an observations/variables table).

Weights: This option is visible only if you selected the observations/variables table format. Activate this option if you want to weight the observations. If you do not activate this option, the weights are considered to be equal to 1. The weights must be greater or equal to 0. If the "Variable labels" option is activated make sure that the header of the selection has also been selected.

Options tab:

Advanced analysis: This option allows you to choose the type of analysis you want to perform on the data. The analysis on supplementary data and the analysis of a subset are only active if the selected data correspond to a contingency table or a more general two-way table. The possible options are:

- **Detrended Correspondence Analysis**: If you select this option you may then enter the parameters that are useful for the calculations, i.e. the number of segments to cut the axes and the number of rescalings to perform. By default, the number of segments is set to 26 and the number of rescalings is set to 4.

- **Supplementary data**: If you select this option you may then enter the number of supplementary rows and/or columns. **Supplementary rows and columns** are passive data that are not taken into account for the computation of the representation space. Their coordinates are computed a posteriori. Notice that supplementary data should be the last rows and/or columns of the data table.

- **Subset analysis**: If you select this option you can then enter the number of **rows and/or columns to exclude** from the subset analysis. See the description section for more information on this topic. Notice that the excluded data should be the last rows and/or columns of the data table.

Non-symmetrical analysis: This option allows computing a Non-Symmetrical Correspondence Analysis, as proposed by Lauro *et al.* (1984).

- **Rows depend on columns**: Select this option if you consider that the row variable depends on the column variable and if you want to analyze the association between both while taking into account this dependency.

- **Columns depend on rows**: Select this option if you consider that the column variable depends on the row variable and if you want to analyze the association between both while taking into account this dependency.

Distance: This option lets you compute a Correspondence Analysis based on the Chi-square distance or on the Hellinger distance as proposed by Rao (1995).

- **Chi-Square**: Select this option to compute classical Correspondence Analysis.
- **Hellinger**: Select this option to compute Correspondence Analysis based on Hellinger distance (HD). This option is not available if the "Non-symmetrical analysis" option has been selected or the case of the Detrended Correspondence Analysis.

To summarize, four approaches of the Correspondence Analysis are proposed:

- Classical Correspondence Analysis (CA): Do not select the "Non-symmetrical analysis" option and select "Chi-Square" distance.
- Non-Symmetrical Correspondence Analysis (NSCA): Select the "Non-symmetrical analysis" option and select "Chi-Square" distance.
- Correspondence Analysis using the Hellinger distance (HD): Do not select the "Non-symmetrical analysis" option and select "Hellinger" distance.
- Detrended Correspondence Analysis (DCA): Select the "Detrended analysis".

Test of independence: Activate this option if you want XLSTAT to compute a test of independence based on the Chi-square statistic.

- **Significance level (%)**: Enter the value of the significance level for the test (default value: 5%).

Filter factors: You can activate one of the two following options in order to reduce the number of factors displayed:

- **Minimum %**: Activate this option and then enter the minimum percentage that should be reached to determine the number of factors to display.
- **Maximum number**: Activate this option to set the maximum number of factors to take into account when displaying the results.

Rotation: Activate this option if you want to apply a rotation to one of the principal coordinate matrices.

- **Number of factors**: Enter the number of factors the rotation is to be applied to.
- **Method**: Choose the rotation method to be used.
- **Based on rows**: Activate this option if you want to apply a rotation from the matrix of the principal coordinates of the rows.
- **Based on columns**: Activate this option if you want to apply a rotation from the matrix of the principal coordinates of the columns.

  Note: If you choose the "Quartimin" method, the choice of the coordinate matrix for the calculations is not possible. In fact, the rotation calculations are performed only on the

principal coordinates of the rows.

- **Kaiser normalization**: Activate this option to apply Kaiser normalization during the rotation calculation.

Missing data tab:

Options for contingency tables and other two-way tables:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Replace missing data by 0: Activate this option if you consider that missing data are equivalent to 0.

Replace missing data by their expected value: Activate this option if you want to replace the missing data by the expected value. The expectation is given by:

$$E(n_{ij}) = \frac{n_{i.}n_{j.}}{n}$$

where $n_{i.}$ is the row sum, $n_{.j}$ is the column sum, and n is the grand total of the table before replacement of the missing data.

Options for the observations/variables tables:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to ignore the observations that contain missing data.

Group missing values into a new category: Activate this option to group missing data into a new category of the corresponding variable.

Outputs tab:

Options specific to the observations/variables tables:

Descriptive statistics: Activate this option to display the descriptive statistics for the two selected variables.

Disjunctive table: Activate this option to display the full disjunctive table that corresponds to the qualitative variables.

Sort categories alphabetically: Activate this option so that the categories of all the variables are sorted alphabetically.

Common options:

Contingency table: Activate this option to display the contingency table.

- **3D view of the contingency table / two-way table**: Activate this option to display the 3D bar chart corresponding to the contingency table or to the two-way table.

Inertia by cell: Activate this option to display the inertia for each cell of the contingency table. This option is not active for the Detrended Correspondence Analysis.

Row and column profiles: Activate this option to display the row and column profiles. This option is not active for the Detrended Correspondence Analysis.

Eigenvalues: Activate this option to display the table and the scree plot of the eigenvalues.

Chi-square (or Hellinger) distances: Activate this option to display the Chi-square (or Hellinger) distances between the row points and between the column points. This option is not active for the Detrended Correspondence Analysis.

Principal coordinates: Activate this option to display the principal coordinates of the row points and the column points.

Standard coordinates: Activate this option to display the standard coordinates of the row points and the column points.

Contributions: Activate this option to display the contributions of the row points and the column points to the principal axes. This option is not active for the Detrended Correspondence Analysis.

Squared cosines: Activate this option to display the squared cosines of the row points and the column points to the principal axes. This option is not active for the Detrended Correspondence Analysis.

Charts tab:

Maps sub-tab:

Options specific to the Detrended analysis:

Detrended analysis plots: Select this option to display the plots for a Detrended Correspondence Analysis.

- **Rows and columns**: Activate this option to display a chart on which the row points and the column points are displayed.
- **Rows**: Activate this option to display a chart on which only the row points are displayed.
- **Columns**: Activate this option to display a chart on which only the column points are displayed.

Labels: Activate this option to display the labels of the categories on the charts.

- **Colored labels**: Activate this option to display the labels with the same color as the corresponding points. If this option is not activated the labels are displayed in black.

Options for other analyses:

Symmetric plots: Activate this option to display the plots for which the row points and the column points play a symmetrical role. These maps are based on the principal coordinates of the row points and the column points.

- **Rows and columns**: Activate this option to display a chart on which the row points and the column points are displayed.
- **Rows**: Activate this option to display a chart on which only the row points are displayed.
- **Columns**: Activate this option to display a chart on which only the column points are displayed.

Asymmetric plots: Activate this option to display the plots for which the row points and the column points play an asymmetrical role. These plots use on the one hand the principal coordinates and on the other hand the standard coordinates.

- **Rows**: Activate this option to display a chart where the row points are displayed using their principal coordinates, and the column points are displayed using their standard coordinates.
- **Columns**: Activate this option to display a chart where the row points are displayed using their standard coordinates, and the column points are displayed using their principal coordinates.
- **Vectors**: Activate this option to display the vectors for the standard coordinates on the asymmetric charts.
- **Lengthening factor**: Activate this option to modulate the length of the vectors.

Contribution biplots: Activate this option to display the contribution biplots for which the row points and the column points play an asymmetrical role. These plots use on the one hand the principal coordinates and on the other hand the contribution coordinates that take into account the weights.

- **Rows**: Activate this option to display a chart where the row points are displayed using their principal coordinates, and the column points are displayed using their contribution coordinates.
- **Columns**: Activate this option to display a chart where the row points are displayed using their contribution coordinates, and the column points are displayed using their principal coordinates.

Confidence ellipses: Activate this option to display confidence ellipses to identify the categories that contribute to the dependency between the row and column categories of the contingency table.

- **Rows**: Active this option to display the confidence ellipses for the rows points on the symmetric plot "Rows".
- **Columns**: Activate this option to display the confidence ellipses for the columns points on the symmetric plot "Columns".

Rotation plots: This option is not active when the "Quartimin" method is selected. Activate this option to display graphs: the biplot with coordinates of rows points and coordinates of columns points calculated after rotation, the graph on which only rows points are displayed, and the graph on which only columns points are displayed. In case the "Based on rows" option is selected, the displayed coordinates of the rows are the principal coordinates and the coordinates of the columns are the standard coordinates.

Row options sub-tab:

These options are not active for the Detrended Correspondence Analysis.

Filter rows: Activate this option to modulate the number of rows displayed:

Options for contingency tables and other two-way tables:

- **Random**: The rows to display are randomly selected. The "Number of rows" N to display must then be specified.
- **N first rows**: The first N rows are displayed on the chart. The "Number of rows" N to display must then be specified.
- **N last rows**: The last N rows are displayed on the chart. The "Number of rows" N to display must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the rows to display.

Common options:

- **Sum(Cos2)>**: Only the rows for which the sum of squared cosines on the given dimensions are larger than a value to enter are displayed on the plots.

Resize row points with Cos2: Activate this option to resize the row points. The sizes of the row points are proportional to the sum of squared cosines on the given dimensions.

Row labels: Activate this option to display the labels of the row categories on the charts.

- **Colored labels**: Activate this option to display the labels with the same color as the corresponding points. If this option is not activated the labels are displayed in black.

Column options sub- tab:

These options are not active for the Detrended Correspondence Analysis.

Filter columns: Activate this option to modulate the number of columns displayed:

Options for contingency tables and other two-way tables:

- **Random**: The columns to display are randomly selected. The "Number of columns" N to display must then be specified.
- **N first columns**: The first N columns are displayed on the chart. The "Number of columns" N to display must then be specified.

- **N last columns**: The last N columns are displayed on the chart. The "Number of columns" N to display must then be specified.

- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the columns to display.

Common options:

- **Sum(Cos2)>**: Only the columns for which the sum of squared cosines on the given dimensions are larger than a value to enter are displayed on the plots.

Resize column points with Cos2: Activate this option to resize the column points. The sizes of the column points are proportional to the sum of squared cosines on the given dimensions.

Column labels: Activate this option to display the labels of the column categories on the charts.

- **Colored labels**: Activate this option to display the labels with the same color as the corresponding points. If this option is not activated the labels are displayed in black.

Results

Descriptive statistics: This table is displayed only if the input data correspond to an observations/variables table. The names of the modalities of each variable, the frequency of each modality and the percentage that each modality represents for the variable are displayed.

Disjunctive table: This table is displayed only if the input data correspond to an observations/variables table. This table is an intermediate table that allows to obtain the contingency table that corresponds to the two selected variables.

Contingency table: The contingency table is displayed at this stage. The **3D bar chart** that follows corresponds to the table.

Inertia by cell: This table displays the inertia that corresponds to each cell of the contingency table.

Test of independence between rows and columns: This test allows us to determine if we can reject the null hypothesis that rows and columns of the table are independent. A detailed interpretation of this test is displayed below the table that summarizes the test statistic.

Eigenvalues and percentages of inertia: The eigenvalues and the corresponding scree plot are displayed. Only the non-trivial eigenvalues are displayed. If a filtering has been requested in the dialog box, it is not applied to this table, but only to the results that follow. In the case of the Detrended Correspondence Analysis, only the eigenvalues are displayed.

Note: The sum of the eigenvalues is equal to the total inertia. To each eigenvalue corresponds a principal axis which accounts for a certain percentage of inertia. This allows us to measure the cumulative percentage of inertia for a given set of dimensions.

A series of results is displayed afterwards, first for the row points, then for the column points:

Weights, distances and squared distances to the origin, inertias and relative inertias: This table gives basic statistics for the points.

Profiles: This table displays the profiles. The mean of the profiles is also displayed except in the case of a Correspondence Analysis based on the Hellinger distance.

Chi-square (or Hellinger) distances: This table displays the Chi- square (or Hellinger) distances between the profile points.

Principal coordinates: This table displays the principal coordinates which are used later to represent projections of the profile points in symmetric and asymmetric plots.

At the end of the rows (or columns) coordinates table, the following button is displayed:  . Click on this button to automatically open the pre-filled dialog box of HAC (Hierarchical Ascending Classification) and perform a classification of the rows (or columns) on the factorial coordinates.

Standard coordinates: This table displays the standard coordinates that are used later to represent projections of unit profile points in asymmetric plots.

Contributions: The contributions are helpful for interpreting the plots. The categories that have most influenced the calculation of the axes are those that have the higher contributions. An approach consists of restricting the interpretation to the categories whose contribution to a given axis is higher than the corresponding relative weight that is displayed in the first column.

Squared cosines: As with other data analysis methods, the analysis of the squared cosines allows us to avoid misinterpretations of the plots that are due to projection effects. If, for a given category, the cosines are low on the axes of interest, then any interpretation of the position of the category is hazardous.

The plots (or maps) are the ultimate goal of Correspondence Analysis, because they allow us to considerably accelerate our understanding of the association patterns in the data table.

Symmetric plots: These plots are exclusively based on the principal coordinates. Depending on the choices made in the dialog box, a symmetric plot mixing row points and column points, a plot with only the row points, and a plot with only the column points, are displayed. The percentage of inertia that corresponds to each axis and the percentage of inertia cumulated over the two axes are displayed on the map. If the "Confidence ellipses" option was selected, confidence ellipses are draw around the points. The confidence ellipses allow the identification of the categories that contribute to the association structure between the variables. The ellipses reflect the information contained in dimensions non-represented on the map.

Asymmetric plots: These plots use the principal coordinates for the rows and the standard coordinates for the columns or vice versa. The percentage of inertia that corresponds to each axis and the percentage of inertia cumulated over the two axes are displayed on the map. In an "Asymmetric row plot", one can study the way the row points are positioned relative to the column vectors. The latter indicate directions: if two row points are displayed in the same direction as a column vector, the row point that is the furthest in the column vector direction is the one that is more associated with the columns.

Contribution biplots: These plots use the principal coordinates for the rows and the contribution coordinates for the columns or vice versa. The percentage of inertia that

corresponds to each axis and the percentage of inertia cumulated over the two axes are displayed on the map. In a "Contribution row plot," one can study the way the row points are positioned relatively to the column vectors while the length of the column vectors take into account their contribution to the building of the biplot.

Note: For the Detrended Correspondence Analysis, the plots displayed are not the same. The "Rows and Columns" option displays the principal coordinates of the rows and the standard coordinates of the columns simultaneously. The "Rows" option displays only the principal coordinates of the rows, while the "Columns" option displays the standard coordinates of the columns. In addition, the percentage of inertia corresponding to each of the axes concerned and the percentage of cumulative inertia of the graph are not displayed. Indeed, the detrending steps cause a distortion of the variance which prevents the interpretation of the eigenvalues as a partition of it. The eigenvalues should therefore be interpreted as indicating the relative importance of each axis.

Where a rotation has been requested, the results of the rotation are displayed with the **rotation matrix** first applied to the row and column coordinates. This is followed by the modified variability percentages associated with each of the axes involved in the rotation. The coordinates, contributions and cosines of the rows and columns after rotation are displayed in the following tables. If the "Rotation Graphs" option has been selected, the graphs representing the coordinates of the rows and columns are displayed.

Example

A tutorial on how to use Correspondence Analysis is available on the XLSTAT Help Center:

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

References

Balbi S. (1997). Graphical displays in non-symmetrical correspondence analysis. In: Blasius J. and Greenacre M. (eds.), Visualisation of Categorical Data. Academic Press, San Diego. pp 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. pp 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., 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. & 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.

Multiple Correspondence Analysis (MCA)

Use this tool to visualize the links between the categories of two or more qualitative variables.

In this section:

Description

Multiple Correspondence Analysis (MCA) is a method that allows studying the association between two or more qualitative variables. MCA is to qualitative variables what Principal Component Analysis is to quantitative variables. Maps can be obtained, where it is possible to visually observe the distances between the categories of the qualitative variables and between the observations.

Multiple Correspondence Analysis (MCA) can also be understood as a generalization of Correspondence Analysis (CA) to the case where there are more than two variables. While it is possible to summarize a table with n observations and p ($p > 2$) qualitative variables in a table whose structure is close to a contingency table, it is much more common in MCA to start from an observations/variables table (for example, from a survey where p questions were submitted to n individuals). XLSTAT also allows the user to start from a full disjunctive table (indicator matrix).

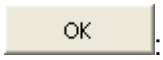
The generation of the disjunctive table is, in any case, a preliminary step of the MCA computations. The p qualitative variables are broken down into p disjunctive tables $Z_1, Z_2, \ldots, Z_p$, composed of as many columns as there are categories in each of the variables. Each time a category c of the j-th variable corresponds to an observation i, one sets the value of $Z_j(i, c)$ to one. The other values of Z_j are zero. The p disjunctive tables are concatenated into a full disjunctive table.

A series of transformations allows the computing of the coordinates of the categories of the qualitative variables, as well as the coordinates of the observations in a representation space that is optimal for a criterion based on inertia. In the case of MCA one can show that the total inertia is equal to the average number of categories minus one. XLSTAT also allows to compute MCA by using the Burt table instead of the disjunctive table. As a matter of fact, the inertia does not only depend on the degree of association between the categories but is seriously inflated. Greenacre (1993) suggested an adjusted version of inertia, inspired from Joint Correspondence Analysis (JCA). This adjustment allows us to have higher and more meaningful percentages for the maps.

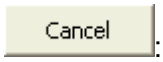
The **analysis of a subset of categories** is a method that has very recently been developed by Greenacre and Pardo (2006). It allows us to concentrate the analysis on some categories only, while still taking into account all the available information in the input table. XLSTAT allows you to select the categories that belong to the subset.

Dialog box

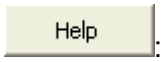
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

The first selection field lets you alternatively select two types of tables:

Observations/variables table: Select this option if your data correspond to a table with n observations described by p qualitative variables. If the headers of the columns have also been selected, make sure the "Variable labels" option is activated.

Disjunctive table: Select this option if your data correspond to a disjunctive table. If the headers of the columns have also been selected, make sure the "Variable labels" option is activated. If this option is activated, supplementary observations and supplementary qualitative variables must also be selected in disjunctive table format.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row contains the variable labels (case of an observations/variables table) or the category labels (case of a disjunctive table).

Weights: Activate this option if you want to weight the observations. If you do not activate this option, the weights are considered to be equal to 1. The weights must be greater or equal to 0. If the "Variable labels" option is activated make sure that the header of the selection has also been selected.

Options tab:

MCA type:

- **Disjunctive table**: If you select this option, MCA will be performed with the observations/variables disjunctive table.
- **Burt table**: If you select this option, MCA will be performed with the observations/variables Burt table.
- **Adjusted inertia**: If you select this option, MCA will be performed with the Burt table and then obtained inertias will be adjusted with the method of Greenacre(1993).

Advanced analysis:

- **None**: If you select this option, MCA will be classically performed on all categories of active variables.
- **Subset analysis**: If you select this option, XLSTAT will ask you to select during the computations the categories that belong to the subset to analyze.

Sort categories alphabetically: Activate this option so that the categories of all the variables are sorted alphabetically.

Variable-Category labels: Activate this option to use variable-category labels when displaying outputs. Variable-Category labels include the variable name as a prefix and the category name as a suffix.

Filter factors: You can activate one of the three following options in order to reduce the number of factors displayed:

- **Minimum %**: Activate this option and then enter the minimum percentage that should be reached to determine the number of factors to display.

- **Maximum number**: Activate this option to set the maximum number of factors to take into account when displaying the results.
- **1/p**: Activate this option to only take into account the factors which eigenvalue is greater than 1/p, where p is the number of variables. This is the default option.

Supplementary data tab:

Supplementary observations: Activate this option if you want to represent additional observations by calculating their coordinates. These observations are neither taken into account for the computation of the correlation matrix, nor for the subsequent calculations (we talk of passive observations as opposed to active observations). If the first row of the data selection for supplementary observations includes a header you must activate the "Variable labels for supp. obs" option. You can also select labels for supplementary observations which will be used for the display.

Supplementary variables: Activate this option if you want to compute a posteriori the coordinates of variables that are not taken into account for the computing of the principal axes (passive variables, as opposed to active variables).

- **Quantitative**: Activate this option if you want to include quantitative supplementary variables. If the headers of the columns of the main table have been selected, you also need to select headers here.
- **Qualitative**: Activate this option if want to include qualitative supplementary variables. If the headers of the columns of the main table have been selected, you also need to select headers here.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to ignore the observations that contain missing data.

Group missing values into a new category: Activate this option to group missing data into a new category of the corresponding variable.

Replace missing data: Activate this option to replace missing data. When a missing data corresponds to a quantitative supplementary variable, they are replaced by the mean of the variable. When a missing data corresponds to a qualitative variable of the initial table (active variables) or to a qualitative supplementary variable (passive variable), a new "Missing" category is create for the variable.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics for the selected variables.

Disjunctive table: Activate this option to display the full disjunctive table that corresponds to the selected qualitative variables.

Burt table: Activate this option to display the Burt table.

3D view of the Burt table: Activate this option to display a 3D visualization of the Burt table.

Eigenvalues: Activate this option to display the table and the scree plot of the eigenvalues.

Display results for:

- **Observations and variables**: Activate this option to display the results that concern the observations and the variables.
- **Observations**: Activate this option to display only the results that concern the observations.
- **Variables**: Activate this option to display only the results that concern the variables.

Principal coordinates: Activate this option to display the principal coordinates.

Standard coordinates: Activate this option to display the standard coordinates.

Contributions: Activate this option to display the contributions.

Squared cosines: Activate this option to display the squared cosines.

Test values: Activate this option to display the test values for the variables.

- **Significance level (%)**: Enter the significance level used to determine if the test values are significant or not.

Charts tab:

Variables sub-tab:

Factorial map of categories: Activate this option to display the chart showing the principal coordinates of categories of active and supplementary qualitative variables.

- **Labels**: Activate this option to show labels of categories on the chart.
- **Colored labels**: Activate this option to show labels in the same color as the points.
- **Color by group**: Activate this option, if you want to color variable points according to levels of a qualitative variable. Then select a vertical series of data that has as many rows

as there are active variables. If headers were selected for the main table, a header must be included in this selection.

- **Resize points with Cos2**: Activate this option so that the categories points sizes are proportional to the respective sum of the squared cosines within the selected subspace.
- **Link categories**: Activate this option so that the categories belonging to a given variable are linked. This option allows to quickly distinguish categories which belong to a variable.

Correlation circle of supplementary variables: Activate this option to display the correlation circle of supplementary quantitative variables.

Filter: Activate this option to modulate the number of variables displayed:

- **Random**: The observations to display are randomly selected. The "Number of variables" N to display must then be specified.
- **N first variables**: The first N variables are displayed on the chart. The "Number of variables" N to display must then be specified.
- **N last variables**: The last N variables are displayed on the chart. The "Number of variables" N to display must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the variables to display.

Observations sub-tab:

Factorial map of observations: Activate this option to display charts representing the principal coordinates of observations.

- **Labels**: Activate this option to have observation labels displayed on the charts. The number of labels displayed can be changed using the filtering option.
- **Colored labels**: Activate this option to show labels in the same color as the points.
- **Color by group**: Activate this option, if you want to color observation points according to levels of a qualitative variable. Then select a vertical series of that that must have as many rows as there are active observations. If headers were selected for the main table, ensure that a label is also present for the variable in this selection.
- **Resize points with Cos2**: Activate this option so that the observation points sizes are proportional to the sum of the corresponding squared cosines within the selected subspace.

Filter: Activate this option to modulate the number of observations displayed:

- **Random**: The observations to display are randomly selected. The "Number of observations" N to display must then be specified.

- **N first rows**: The first N observations are displayed on the chart. The "Number of observations" N to display must then be specified.
- **N last rows**: The last N observations are displayed on the chart. The "Number of observations" N to display must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to display.
- **Sum(Cos2)>**: Only the observations for which the sum of squared cosines (communalities) are bigger than a value to enter are displayed on the plots.

Biplots sub-tab:

Biplots: Activate this option to display charts representing the observations and categories simultaneously.

- **Symmetric plot**: Activate this option to display on the same chart principal coordinates of observations and principal coordinates of categories.
- **Asymmetric plots**: Activate this option to display asymmetric charts. These charts use principal coordinates and standardized coordinates. Two kinds of charts are available:
- **Asymmetric row plot**: Activate this option to display on the same chart principal coordinates of observations and standardized coordinates of categories.
- **Asymetric column plot**: Activate this option to display on the same chart principal coordinates of categories and standardized coordinates of observations.

Options for variables:

- **Categories labels**: Activate this option to have observation labels displayed on the biplots.
- **Vectors**: Activate this option to display the initial variables in the form of vectors.
- **Supplementary variables**: If you have included supplementary variables in the PCA, activate this option to display them on the biplot.
- **Filter variables**: If you used a filter variable to display variables, the same filter variable will be used to filter the display of variables on the biplot.

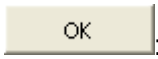
Options for observations:

- **Observations labels**: Activate this option to have observation labels displayed on the biplots.
- **Vectors**: Activate this option to display the observations in the form of vectors.

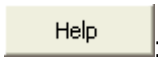
- **Supplementary observations**: If you have included supplementary observations in the MCA, activate this option to display them on the biplot.
- **Filter variables**: If you used a filter variable to display observations, the same filter variable will be used to filter the display of observations on the biplot.

Dialog box (subset categories)

This dialog is displayed if you selected the **Advanced analysis** / **Subset analysis** option in the MCA dialog box.



: Click this button to start the computations.



: Click this button to display the help.

The **list of categories** that corresponds to the complete set of active qualitative variables is displayed so that you can select the subset of categories on which the analysis will be focused.

All: Click this button to select all the categories.

None: Click this button to deselect all the categories.

Results

Descriptive statistics: This table is displayed only if the input data correspond to an observations/variables table.

Disjunctive table: This table is displayed only if the input data correspond to an observations/variables table. This table is an intermediary table that allows us to obtain the contingency table that corresponds to the two selected variables.

Burt table: The Burt table is displayed only if the corresponding option is activated in the dialog box. The **3D bar chart** that follows is the graphical visualization of this table.

Eigenvalues and percentages of inertia: The eigenvalues, the percentages of inertia and the corresponding scree plot are displayed. Only the non- trivial eigenvalues are displayed. If a filtering has been requested in the dialog box, it is not applied to this table, but only to the results that follow.

A series of results is displayed afterwards, first for the variables, then for the observations:

Principal coordinates: This table displays the principal coordinates which are used later to represent projections of profile points in symmetric and asymmetric plots. At the end of the observations coordinates table, the following button is displayed:  . Click on this button to

automatically open the pre-filled dialog box of HAC (Hierarchical Ascending Classification) and perform a classification of the observations on the factorial coordinates.

Standard coordinates: This table displays the standard coordinates which are used later to represent projections of unit profile points in asymmetric plots.

Contributions: The contributions are helpful for interpreting the plots. The categories that have influenced the most the calculation of the axes are those that have the higher contributions. A shortcut consists of restricting the analysis to the categories which contribution on a given axis is higher than the corresponding relative weight that is displayed in the first column.

Axes homogeneity index: This index developed by our team is very useful to determine if the contributions of the observations are homogeneous for the different axes. It is constructed as the proportion of observations with an absolute contribution > 1/n. An index above 0.4 indicates a very good homogeneity with well represented observations. On the other hand, an index lower than 0.1 should be a warning to the user who should check if there are no outliers in the variables constructing the axis that would distort its interpretation (the outliers would then be the observations that stand out from the others on the axis in question).

Squared cosines: As with other data analysis methods, the analysis of the squared cosines allows us to avoid misinterpretations of the plots that are due to projection effects. If, for a given category, the cosines are low on the axes of interest, then any interpretation of the position of the category is hazardous.

The plots (or maps) are the ultimate goal of Multiple Correspondence Analysis, because they considerably facilitate our interpretation of the data.

Symmetric plots: These plots are exclusively based on the principal coordinates. Depending on the choices made in the dialog box, a symmetric plot mixing observations and variables, a plot showing only the categories of the variables, and a plot showing only the observations, are displayed. The percentage of adjusted inertia that corresponds to each axis and the percentage of adjusted inertia cumulated over the two axes are displayed on the map.

Asymmetric plots: These plots use the principal coordinates for the categories of the variables and the standard coordinates for the observations and vice versa. The percentage of inertia that corresponds to each axis and the percentage of inertia cumulated over the two axes are displayed on the map. On an "asymmetric observations plot", on can study the way the observations are positioned relatively to the category vectors. The later indicate directions: if two observations are displayed in the same direction as a category vector, the observation that is the furthest in the category vector direction is more likely to have selected that category of response.

Example

A tutorial on how to use Multiple Correspondence Analysis is available on the XLSTAT Help Center:

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

References

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., Pardo R. (2006). Multiple correspondence analysis of subsets of response categories. In Multiple Correspondence Analysis and Related Methods (eds Michael Greenacre & 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)

Use multidimensional scaling to represent in a two- or three-dimensional space the observations for which only a proximity matrix (similarity or dissimilarity) is available.

In this section:

Description

Multidimensional Scaling (MDS) is used to go from a **proximity** matrix (similarity or dissimilarity) between a series of N objects to the coordinates of these same objects in a p-dimensional space. p is generally fixed at 2 or 3 so that the objects may be visualized easily. For example, with MDS, it is possible to reconstitute the position of towns on a map very precisely from the distances in kilometers (the dissimilarity in this case being the Euclidean distance) between the towns, modulo a rotation and a symmetry.

This example is only intended to demonstrate the performance of the method and to give a general understanding of how it is used. Practically, MDS is often used in psychometry (perception analysis) and marketing (distances between products obtained from consumer classifications) but there are applications in a large number of domains.

If the starting matrix is a similarity matrix (a similarity is greater the nearer the objects are), it will automatically be converted into a dissimilarity matrix for the calculations. The conversion is carried out by subtracting the matrix data from the value of the diagonal.

There are two types of MDS depending on the nature of the dissimilarity observed:

- **Metric MDS**: The dissimilarities are considered as continuous and giving exact information to be reproduced as closely as possible. There are a number of sub-models:
- Absolute MDS: the distances obtained in the representation space must correspond as closely as possible to the distances observed in the starting dissimilarity matrix.
- Ratio MDS: the distances obtained in the representation space must correspond as closely as possible to the distances observed in the initial matrix using a near proportionality factor (the factor being identical for all pairs of distances).
- Interval MDS: the distances obtained in the representation space must correspond as closely as possible to the distances observed in the initial matrix using a near linear relationship (the linear relationship being identical for all pairs of distances).

- Polynomial MDS: the distances obtained in the representation space must correspond as closely as possible to the distances observed in the initial matrix using a near 2-nd-degree polynomial relationship (the polynomial relationship being identical for all pairs of distances).

Note: the absolute model is used to compare distances in the representation space with those in the initial space. The other models have the advantage of speeding up the calculations.

- **Non metric MDS**: with this type of MDS, only the order of the dissimilarities counts. In other words, the MDS algorithm does not have to try to reproduce the dissimilarities but only their order. Two models are available:

- Ordinal (1): the order of the distances in the representation space must correspond to the order of the corresponding dissimilarities. If there are two dissimilarities of the same rank, then there are no restrictions on the corresponding distances. In other words, dissimilarities of the same rank need not necessarily give equal distances in the representation space.

- Ordinal (2): the order of the distances in the representation space must correspond to the order of the corresponding dissimilarities. If dissimilarities exist in the same rank, the corresponding distances must be equal.

The MDS algorithms aim to reduce the difference between the disparity matrix from the models and the distance matrix obtained in the representation configuration. For the absolute model, the disparity is equal to the dissimilarity of the starting matrix. The difference is measured through the Stress, several variations of which have been proposed:

- Raw Stress:

$$\sigma_r = \sum_{i<j} w_{ij}(D_{ij} - d_{ij})^2,$$

where D_{ij} is the disparity between individuals i and j, d_{ij} is the Euclidean distance on the representation for the same individuals, and w_{ij} is the weight of the ij proximity (value is 1 by default).

- Normalized Stress:

$$\sigma_r = \frac{\sum_{i<j} w_{ij}(D_{ij} - d_{ij})^2}{\sum_{i<j} w_{ij} D_{ij}^2}.$$

- Kruskal's stress 1:

$$\sigma_r = \sqrt{\frac{\sum_{i<j} w_{ij}(D_{ij} - d_{ij})^2}{\sum_{i<j} w_{ij} d_{ij}^2}}.$$

- Kruskal's stress 2:

$$\sigma_r = \sqrt{\frac{\sum_{i<j} w_{ij}(D_{ij} - d_{ij})^2}{\sum_{i<j} w_{ij}(d_{ij} - \bar{d})^2}}.$$

where $\bar{d}$ is the average of the distances on the representation.

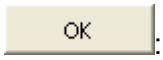
Note: for a given number of dimensions, the weaker the stress, the better the quality of the representation. Furthermore, the higher the number of dimensions, the weaker the stress.

To find out whether the result obtained is satisfactory and to determine which is the correct number of dimensions needed to give a faithful representation of the data, the evolution in the stress with the number of dimensions and the point from which the stress stabilizes may be observed. The Shepard diagram is used to observe any ruptures in the ordination of the distances. The more the chart looks linear, the better the representation. For the absolute model, for an ideal representation, the points must be aligned along the first bisector.

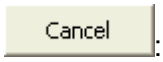
There are several MDS algorithms including, in particular, ALSCAL (Takane *et al*. 1977) and SMACOF (*Scaling by MAjorizing a COnvex Function*) which minimizes the "Normalized Stress" (de Leeuw, 1977). XLSTAT uses the SMACOF algorithm.

Dialog box

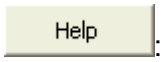
The dialog box is made up of several tabs corresponding to the various options for controlling the calculations and displaying the results. A description of the various components of the dialog box are given below.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.



: Click this button to display help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. if the arrow points down XLSTAT allows you to select data by columns or by range. If the arrow points to the right, XLSTAT allows you to select data by rows or by range.

General tab:

The main data entry field is used to select one of two types of table:

Data: Select a similarity or dissimilarity matrix. If only the lower or upper triangle is available, the table is accepted. If differences are detected between the lower and upper parts of the selected matrix, XLSTAT warns you and offers to change the data (by calculating the average of the two parts) to continue with the calculations.

Dissimilarities / Similarities: Choose the option that corresponds to the type of your data.

Model: Select the model to be used. See description for more details.

Dimensions: Enter the minimum and maximum number of dimensions for the object representation space. The algorithm will be repeated for all dimensions between the two boundaries.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet in the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if you have included row and column labels in the selection.

Weights: Activate this option if the data are weighted. You then select a weighting matrix (without selecting labels for rows and columns). If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0.

Options tab:

Stress: Choose the type of stress to be used for returning the results, given that the SMACOF algorithm minimizes the raw stress. See the description section for more details.

Initial configuration:

- **Random**: Activate this option to make XLSTAT generate the starting configuration randomly. Then enter the number of times the algorithm is to be repeated from a new randomly-generated configuration. The default value for the number of repetitions is 100. Note: the configuration displayed in the results is the repetition for which the best result was found.
- **User defined**: Activate this option to select an initial configuration which the algorithm will use as a basis for carrying out optimization.

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the SMACOF algorithm. Stress Optimization is stopped when the maximum number of iterations has been exceeded. Default value: 100.
- **Convergence**: Enter the minimum value of evolution in stress from one iteration to another which, when reached, means that the algorithms is considered to have converged. Default value: 0.00001.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Ignore missing data: If you activate this option, XLSTAT does not include proximities corresponding to missing data when minimizing stress.

Outputs tab:

Distances: Activate this option to display the matrix of Euclidean distances corresponding to the optimum configuration.

Disparities: Activate this option to display the disparity matrix corresponding to the optimum configuration.

Residual distances: Activate this option to display the matrix of residual distances corresponding to the difference between the distance matrix and the disparity matrix.

Charts tab:

Evolution of stress: Activate this option to display the stress evolution chart according to the number of dimensions in the configuration.

Configuration: Activate this option to display the configuration representation chart. This chart is only displayed for the configuration in a two-dimensional space if this has been calculated.

- **Labels**: Activate this option if you want object labels to be displayed.
- **Colored labels**: Activate this option to show labels in the same color as the points.
- **Shepard diagram**: Activate this option to display the Shepard diagram.

Results

Stress after minimization: This table shows the final stress obtained, the number of iterations required and the level of convergence reached for the dimensions considered. Where multiple dimensions were considered, a chart is displayed showing the stress evolution as a function of the number of dimensions.

The results which follow are displayed for each of the dimensions considered.

Configuration: This table shows the coordinates of objects in the representation space. If this is a two-dimensional space, a graphic representation of the configuration is provided. If you have XLSTAT-3DPlot, you can also display a three-dimensional configuration.

Distances measured in the representation space: This table shows the distances between objects in the representation space.

Disparities computed using the model: This table shows the disparities calculated according to the model chosen (absolute, interval, etc.).

Residual distances: These distances are the difference between the dissimilarities of the starting matrix and the distances measured in the representation space.

Comparative table: This table is used to compare dissimilarities, disparities and distances and the ranks of these three measurements for all paired combinations of objects.

Shepard diagram: This chart compares the disparities and the distances to the dissimilarities. For a metric model, the representation is better the more the points are aligned with the first bisector of the plan. For a non-metric model, the model is better the more regularly the line of dissimilarities/disparities increases. Furthermore, the performance of the model can be evaluated by observing if the (dissimilarity/distance) points are near to the (dissimilarity/disparity) points.

Example

A tutorial on how to use Multidimensional Scaling is available on the XLSTAT Help Center:

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

References

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.

k-means clustering

Use k-means clustering to make up homogeneous groups of objects (clusters) on the basis of their description by a set of quantitative variables.

In this section:

Description

k-means clustering was introduced by McQueen in 1967. Other similar algorithms had been developed by Forgey (1965) (moving centers) and Friedman (1967).

k-means clustering has the following advantages:

An object may be assigned to a cluster during one iteration then change cluster in the following iteration, which is not possible with Agglomerative Hierarchical Clustering, where assignment is irreversible.

By multiplying the starting points and the repetitions, several solutions may be explored.

The disadvantage of this method is that it does not give a consistent number of clusters or enable the proximity between clusters or objects to be determined.

The k-means and AHC methods are therefore complementary.

Note: if you want to take qualitative variables into account in the clustering, you must first perform a Multiple Correspondence Analysis (MCA) and consider the resulting coordinates of the observations on the factorial axes as new variables.

Principle of the k-means method

k-means clustering is an iterative method which, wherever it starts from, converges on a solution. The solution obtained is not necessarily the same for all starting points. For this reason, the calculations are generally repeated several times in order to choose the optimal solution for the selected criterion.

For the first iteration, a starting point is chosen which consists of associating the center of the k clusters with k objects (either taken at random or not). Afterwards, the distance between the

objects and the k centers are calculated, and the objects are assigned to the centers they are nearest to. Then the centers are redefined from the objects assigned to the various clusters. The objects are then reassigned depending on their distances from the new centers. And so on until convergence is reached.

Classification criteria

Several classification criteria may be used to reach a solution. XLSTAT offers four criteria for the k-means minimization algorithm.

Trace(W): The W trace, *pooled SSCP matrix*, is the most traditional criterion. Minimizing the W trace for a given number of clusters amounts to minimizing the total within-cluster variance — in other words, minimizing the heterogeneity of the groups. This criterion is sensitive to effects of scale. In order to avoid giving more weight to certain variables and not to others, the data must be normalized beforehand. Moreover, this criterion tends to produce clusters of the same size.

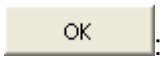
Determinant(W): The determinant of W, *pooled within covariance matrix*, is a criterion considerably less sensitive to effects of scale than the W trace criterion. Furthermore, group sizes may be less homogeneous than with the trace criterion.

Wilks lambda: The results given by minimizing this criterion are identical to that given by the determinant of W. Wilks' lambda criterion corresponds to the division of determinant(W) by determinant(T) where T is the total inertia matrix. Dividing by the determinant of T always gives a criterion between 0 and 1.

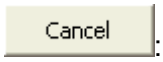
Trace(W) / Median: If this criterion is chosen, the cluster centroid is not the mean point of the cluster but the median point, which corresponds to an object of the cluster. The use of this criterion gives rise to longer calculations.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 ![Select_list.png][image-9]{: width="26" height="25"} ![Select_file.png][image-10]{: width="26" height="25"} : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files ![Select_file_choosefile.png][image-11]{: width="26" height="25"}.

General tab:

Observations/variables table: Select a table comprising N objects described by P descriptors. If column headers have been selected, check that the "Variable labels" option has been activated.

Dissimilarity Index: Choose a distance among the three available on XLSTAT. The Euclidean distance is the most used distance for k-means and is applied in most cases. The cosine dissimilarity is recommended in order to analyze textual data. The Jaccard index is recommended for datasets that require a fine analysis.

Classification criterion: Choose the classification criterion (see the description section for more details).

Number of clusters: Enter the number of clusters to be created by the algorithm. You can let the number of clusters vary between two bounds.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selections (Observations/variables table, row labels, row weights, column weights) contains a label.

Row labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Column labels" option is activated, you need to include a header in the selection. If this option is not activated, the observation labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Column weights: Activate this option if the columns are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, make sure the "Column labels" option is activated.

Row weights: Activate this option if the rows are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header

has been selected, make sure the "Column labels" option is activated.

Options tab:

Cluster rows: Activate this option if you want to create clusters of objects in rows described by descriptors in columns.

Cluster columns: Activate this option if you want to create clusters of objects in columns described by descriptors in rows.

Center: Activate this option if you want to center the data before starting the calculations.

Reduce: Activate this option if you want to reduce the data before starting the calculations.

You can then select whether you want to apply the transformation on the rows or the columns.

Results in the original space: Activate this option to display the results in the original space. If the center/reduce options are activated and this option is not activated, the results are provided in the standardized space.

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the k-means algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 500.
- **Convergence**: Enter the minimum value of evolution for the chosen criterion from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.00001.

Initial partition: Use these options to choose the way the first partition is chosen, in other words, the way objects are assigned to clusters in the first iteration of the clustering algorithm.

- N **clusters by data order**: Objects are assigned to clusters depending on their order.
- **Random**: Objects are assigned to clusters randomly.
- **User defined**: Objects are assigned to clusters according to an indicator variable defined by the user. The user must in this case select a column indicator variable containing as many rows as objects (with an optional header), and the clusters must be defined by the values 1 to k where k is the number of clusters. If the "Column labels" option is activated, you need to include a header in the selection.
- **Defined by centers**: The user has to select the k centers corresponding to the k clusters. In cases where the clustering is done on rows, the number of rows defines the number of clusters and the number of columns must be the same as the number of columns of the data table. On the other hand, if the clustering is done on columns, the number of columns defines the number of clusters and the number of rows must be the same as the number of rows in the data table. If the "Column labels" option is activated, the first cell of the selection must contain a header.
- **K++**: This option lets you define centers according to k-means++ algorithm introduced by Rafail Ostrovsky, Yuval Rabani, Leonard Schulman and Chaitanya Swamy in 2006. The

first center is chosen randomly among the observations. The next is chosen among the observations depending on the distance between the observation and the center. The further the observation is from the center the higher the probability it will be chosen. The $k-2$ remaining centers are chosen according to the same method. This method allows you to start with centers chosen evenly in the dataset which generally increase the quality of the partition and the speed at which the algorithm reaches the solution. But this algorithm take times to compute and with large and complex datasets (with a lot of centers), it is recommended to use K|| algorithm.

- **K||**: This option lets you define centers according to K|| or to the "Scalable K-means" algorithm introduced by Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar and Sergei Vassilvitskii in 2012. It is derived from the K++ algorithm and lets you choose the centers with parallelization. Like K++, The first center is chosen randomly among the observations, but at each iteration of the algorithm, $\tilde{k}/2$ observations are chosen randomly and independently with the same method as K++. After a number of runs, depending on the size of the dataset, the X centers obtained are then reclustered into k centers using K++. This algorithm has the advantage of being much faster than K++ for two reasons : the first step is mainly decreasing the amount of relevant observations to be processed by K++ and the independent choice of the centers for each iteration allows for a parallel implementation of the first step.

Prediction tab:

Prediction: Activate this option if you want to select data to use in prediction mode. If you activate this option, you need to make sure that the prediction dataset is structured like the estimation dataset: the same variables with the same order in the selections.

Observations/variables table: Select a table comprising new objects described by the same P descriptors as the learning table. If column headers have been selected, check that the "Variable labels" option has been activated.

Observation labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Variable Labels" option is activated, you need to include a header in the selection. If this option is not activated, the observation labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Variable labels: Activate this option if the first row of the data selections (variables and observation labels) includes a header.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected for each activated dataset (training, prediction).

Correlation matrix: Activate this option to display a view of the correlations between the various variables selected.

Optimization summary: Activate this option to display the optimization summary.

Centroids: Activate this option to display the table of centroids of the clusters.

Central objects: Activate this option to display the coordinates of the nearest object to the centroid for each cluster.

Contribution (Analysis of variance) : Activate this option to display a table giving the contribution of each variable.

Results by cluster: Activate this option to display a table giving the statistics and the objects for each of the clusters.

Results by object: Activate this option to display a table giving the cluster each object is assigned to and its distance from the centroids of its cluster (in the initial object order).

- **Correlation with centroids**: Activate this option to display the Pearson correlation between an object and the centroid of its cluster.
    - **Noisy observation**: Activate this option to display a column indicating which observation is noisy. An observation is noisy if the correlation with its cluster centroid is smaller than the threshold you choose.
- **Silhouette score**: Activate this option to display the silhouette score of each object.
    - **Mean by cluster**: Activate this option to display a table giving the mean silhouette score for each of the clusters.

Charts tab:

Evolution of the criterion: Activate this option for the evolution chart of the chosen criterion.

Profile plot: Activate this option to display a plot that allows you to compare the means of the different clusters that have been created.

Silhouette score: Activate this option to display a plot showing silhouette score of each object.

Silhouette score (Means): Activate this option to display a plot showing the mean silhouette score of each cluster.

Results

Summary statistics: This table displays the descriptors of the objects, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation.

Correlation matrix: This table is displayed to give you a view of the correlations between the various variables selected.

Evolution of the within-cluster inertia: If you have selected a number of clusters between two bounds, XLSTAT displays at first the evolution of the within-cluster inertia, which reduces mathematically when the number of clusters increases. If the data is distributed homogeneously, the decrease is linear. If there is actually a group structure, an elbow is observed for the relevant number of clusters.

Evolution of the silhouette score: If you have selected a number of clusters between two bounds, a table with its associated chart shows the evolution of the silhouette score for each k. The optimal number of clusters is the k whose silhouette score is closest to 1.

Optimization summary: This table shows the evolution of the within-cluster variance. If several repetitions have been requested, the results for each repetition are displayed. The repetition giving the best classification is displayed in bold.

Statistics for each iteration: This table shows the evolution of miscellaneous statistics calculated as the iterations for the repetition proceed, given the optimum result for the chosen criterion. If the corresponding option is activated in the Charts tab, a chart showing the evolution of the chosen criterion as the iterations proceed is displayed.

Note: if the values are standardized (option in the Options tab), the results for the optimization summary and the statistics for each iteration are calculated in the standardized space. On the other hand, the following results are displayed in the original space if the "Results in the original space" option is activated.

Inertia decomposition for the optimal classification: This table shows the within-cluster inertia, the between-cluster inertia and the total inertia.

Initial cluster centroids: This table shows the initial cluster centroids computed thanks to the initial random partition or with K|| and K++ algorithms. In case you defined the centers, this table shows the selected cluster centroids.

cluster centroids: This table shows the cluster centroids for the various descriptors.

Distance between the cluster centroids: This table shows the distances between the cluster centroids for the various descriptors.

Central objects: This table shows the coordinates of the nearest object to the centroid for each cluster.

Distance between the central objects: This table shows the distances between the cluster central objects for the various descriptors.

Results by cluster: The descriptive statistics for the clusters (number of objects, sum of weights, within-cluster variance, minimum distance to the centroid, maximum distance to the centroid, mean distance to the centroid) are displayed in the first part of the table. The second part shows the objects.

Results by object: This table shows the assignment cluster for each object in the initial object order. * **Distance to centroid**: This column shows the distance between an object and its cluster centroids.
* **Correlations with centroids**: This column shows the Pearson correlation between an object and its cluster centroids. * **Noisy observation**: This column indicates which observation is noisy with a bold "Yes" displayed. * **Silhouette scores**: This column shows the silhouette score of each object.

The silhouette score measures the quality of the classification of an observation in a cluster. It is formulated as: $$ s(i) = \frac{b(i) - a(i)}{max(a(i),b(i))}, $$ where $a(i)$ is the mean distance between i and all other points in the same cluster, and $b(i)$ is the mean distance between i and all other points in the nearest cluster. The silhouette score varies between -1 and 1 and the closer its value is to 1 the better an observation lies within its cluster.

Silhouette scores (Mean by cluster): This table and its graph are displayed and show the mean silhouette score of each cluster and the silhouette score for the optimal classification (mean of means by cluster).

Contribution (Analysis of variance) : This table indicates the variables that contribute the most to the separation of the clusters by performing an ANOVA.

Profile plot: This chart allows you to compare the means of the different clusters that have been created.

Example

A tutorial on k-means clustering is available on the XLSTAT Help Center:

http://www.xlstat.com/demo-cluster2.htm

References

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 castigation. *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 castigation 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.

Agglomerative Hierarchical Clustering (AHC)

Use Agglomerative Hierarchical Clustering to make up homogeneous groups of objects (clusters) on the basis of their description by a set of variables, or from a matrix describing the similarity or dissimilarity between the objects.

In this section:

Description

Principle of AHC

Agglomerative Hierarchical Clustering (AHC) is an iterative classification method whose principle is simple.

The process starts by calculating the dissimilarity between the N objects. Then two objects, which when clustered together minimize a given agglomeration criterion, are clustered together thus creating a cluster comprising these two objects. Then the dissimilarity between this cluster and the N-2 other objects is calculated using the agglomeration criterion. The two objects or clusters of objects whose clustering together minimizes the agglomeration criterion are then clustered together. This process continues until all the objects have been clustered.

These successive clustering operations produce a binary clustering tree (dendrogram), whose top node is the cluster that contains all the observations. This dendrogram represents a hierarchy of partitions.

It is then possible to choose a partition by truncating the tree at a given level, with the level depending upon either user-defined constraints (the user knows how many clusters are to be obtained) or more objective criteria.

Similarities and dissimilarities

The proximity between two objects is measured by measuring at what point they are similar (similarity) or dissimilar (dissimilarity). If the user chooses a similarity, XLSTAT converts it into a dissimilarity as the AHC algorithm uses dissimilarities. The conversion for each object pair consists in taking the maximum similarity for all pairs and subtracting from this the similarity of the pair in question.

The **similarity** coefficients proposed are as follows: Cooccurrence, Cosine, Covariance (n-1), Covariance (n), Dice coefficient (also known as Sorensen coefficient), General similarity, Gower coefficient, Inertia, Jaccard coefficient, Kendall correlation coefficient, Kulczinski coefficient, Ochiai coefficient, Pearson's correlation coefficient, Pearson Phi, Percent agreement, Rogers & Tanimoto coefficient, Sokal & Michener coefficient (or simple matching coefficient), Sokal & Sneath coefficient (1), Sokal & Sneath coefficient (2), Spearman correlation coefficient, Squared correlations.

Here are the **dissimilarity** coefficients proposed: Bhattacharya's distance, Bray and Curtis' distance, Canberra's distance, Chebychev's distance, Chi² distance, Chi² metric, Chord distance, Squared chord distance, Dice coefficient, Euclidean distance, Geodesic distance, Jaccard coefficient, Kendall dissimilarity, Kulczinski coefficient, Mahalanobis distance, Manhattan distance, Ochiai coefficient, Pearson's dissimilarity, Pearson's Phi, General dissimilarity, Rogers & Tanimoto coefficient, Sokal & Michener's coefficient, Sokal & Sneath's coefficient (1), Sokal & Sneath coefficient (2), Spearman dissimilarity, Squared correlations.

Note: some of the abovementioned coefficients should be used with binary data only.

Quality of a hierarchical clustering

The quality of an agglomerative hierarchical clustering might be assessed by the cophenetic correlation coefficient. It uses a measure of distance called the cophenetic distance that can be estimated on the dendrogram obtained from a classification.

The cophenetic distance separating 2 observations is given by the height of the dendrogram at which the 2 observations belong to the same cluster.

Finally, the cophenetic correlation is estimated as the Pearson correlation coefficient between the dissimilarity matrix used for the AHC and the cophenetic distance matrix. The closer to 1 the correlation, the better the quality of the clustering.

Agglomeration methods

To calculate the dissimilarity between two groups of objects A and B, different strategies are possible. XLSTAT offers the following methods:

Simple linkage: The dissimilarity between A and B is the dissimilarity between the object of A and the object of B that are the most similar. Agglomeration using simple linkage tends to contract the data space and to flatten the levels of each step in the dendrogram. As the dissimilarity between two elements of A and of B is sufficient to link A and B, this criterion can lead to very long clusters (chaining effect) while they are not homogeneous.

Complete linkage: The dissimilarity between A and B is the largest dissimilarity between an object of A and an object of B. Agglomeration using complete linkage tends to dilate the data space and to produce compact clusters.

Unweighted pair-group average linkage: The dissimilarity between A and B is the average of the dissimilarities between the objects of A and the objects of B. Agglomeration using Unweighted pair-group average linkage is a good compromise between the two preceding criteria, and provides a fair representation of the data space properties.

Weighted pair-group average linkage: The average dissimilarity between the objects of A and of B is calculated as the sum of the weighted dissimilarities, so that equal weights are assigned to both groups. As with unweighted pair-group average linkage, this criterion provides a fairly good representation of the data space properties.

Flexible linkage: This criterion uses the parameter $\hat{a}$ that varies between [-1,+1]; this can generate a family of agglomeration criteria. For $\hat{a} = 0$ the criterion is weighted pair-group average linkage. When $\hat{a}$ is near to 1, chain- like clusters result, but as $\hat{a}$ decreases and becomes negative, more and more dilatation is obtained.

Ward's method: This method is the most used one, it aggregates two groups so that within-group inertia increases as little as possible to keep the clusters homogeneous. This criterion, proposed by Ward (1963), can only be used in cases with quadratic distances, i.e. cases of Euclidean distance and Chi-square distance.

Truncate the dendrogram

Cutting or truncating the dendrogram allows you to obtain a partition after having performed an agglomerative hierarchical clustering. XLSTAT offers 7 methods to cut the dendrogram, including 5 methods that cut the dendrogram into a partition according to a specific criterion.

Three widely used indices are proposed: the Hartigan index, the Silhouette coefficient and the Calinski and Harabasz index. First, XLSTAT lets you enter an interval of number of classes k within which the indices will recommend the number of classes to choose. Then, for each number of classes included in this interval, a cut is made in the dendrogram to create a partition in k classes. Finally, the indices are computed thanks to the obtained partition:

- *Hartigan index*: $$ H(k) = (\frac{W_k}{W_{k+1}}-1)(n-k-1), $$ where n is the number of objects, k the number of clusters, W_k and W_{k+1} are the within-cluster sum of squares for the partition with k clusters and $k+1$ clusters.

- *Silhouette score*: $$ S(k) = \frac{1}{n}\sum_ { i = 1} ^ { n }\frac{b(i) - a(i)}{max(a(i),b(i))}, $$

where $a(i)$ is the mean distance between i and all other points in the same cluster, and $b(i)$ is the mean distance between i and all other points in the nearest cluster.

- *Calinski and Harabasz index*: $$ CH(k) = \frac{B_k(n-k)}{W_k(k-1)} , $$ where B_k between-cluster sum of squares for the partition with k clusters.

These indices help you automatically choose the best number of clusters k. If you choose the Silhouette score or the Calinski and Harabasz index to truncate the dendrogram, the best k is the k giving the greater $S(k)$ or the greater $CH(k)$. If you choose the Hartigan index, the best k is the k giving the greater $H(k-1) - H(k)$.

Note: Our teams have adapted the Hartigan index and the Calinski and Harabasz index so that you can use them if you have not chosen the Euclidean distance and/or the Ward criteria.

Two other methods let you cut the dendrogram through the evolution of the entropy or the inertia between each level. The dendrogram will be cut between the 2 levels giving the greater evolution.

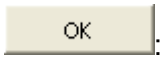
Finally, you can define the level (cut-off) by hand or define the number of clusters you want.

Consolidation using k-means

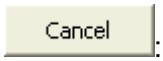
It is possible to perform the k-means algorithm over an Agglomerative Hierarchical Clustering (AHC) with the Euclidean distance or the Jaccard distance. Indeed, k-means is performed using the partition generated after the truncation as an initial partition of the algorithm. This method allows you to get a clustering with a lower within-cluster variation (or equal in the k-means algorithm did not find a better partition). In other words, the consecutive use of the two algorithms, AHC and k-means, results in a better quality clustering.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options, ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 ![Select_list.png][image-9]{: width="26" height="25"} ![Select_file.png][image-10]{: width="26" height="25"} : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files ![Select_file_choosefile.png][image-11]{: width="26" height="25"}.

General tab:

Observations/variables table/ Proximity matrix: Choose the option that corresponds to the format of your data, then select the data. For the **Observations/variables table** option, select a table comprising N objects described by P quantitative descriptors. For a **Proximity matrix**, select a squared matrix giving the proximities between the objects. If column headers have

been selected, check that the "Variable labels" option has been activated. For a proximity matrix, if column labels have been selected, row labels must also be selected.

Proximity type: **similarities / dissimilarities**: Choose the proximity type to be used. The data type and proximity type determine the list of possible indexes for calculating the proximity matrix.

Agglomeration method: Choose the agglomeration method (see the description section for more details).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet in the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selections (Observations/variables table, row labels, row weights, column weights) contains a label. Where the selection is a proximity matrix, if this option is activated, the first column must also include the object labels.

Row labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Column labels" option is activated, you need to include a header in the selection. If this option is not activated, the observation labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Row weights: Activate this option if the rows are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column labels" option is activated.

Column weights: Activate this option if the columns are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column labels" option is activated.

Options tab:

Cluster rows: Activate this option if you want to create clusters of objects in rows described by data in columns.

Cluster columns: Activate this option if you want to create clusters of objects in columns described by data in rows.

Center: Activate this option if you want to center the data before starting the calculations.

Reduce: Activate this option if you want to reduce the data before starting the calculations.

You can then select whether you want to apply the transformation on the rows or the columns.

Results in the original space: Activate this option to display the results in the original space. If the center/reduce options are activated and this option is not activated, the results are provided in the standardized space.

Within-cluster variances: Activate this option to select the within-cluster variances. This option is only active if object weights have been selected (row weights if you are clustering rows, column weights if you are clustering columns). This option can be used if you previously clustered the objects using another method (k-means for example) and want to use a method such as **unweighted pair group averages** to cluster the groups previously obtained. If a column header has been selected, check that the "Column labels" option is activated.

Truncation: Activate this option if you want XLSTAT to automatically define the truncation level, and therefore the number of clusters to retain using one of these 5 methods: **Hartigan index**, **Silhouette index**, **Calinski and Harabasz index**, **Entropy** or **Inertia**. You can also define the **number of clusters** to create (you can let the number of clusters vary between two bounds), or the **level** at which the dendrogram is to be truncated.

Consolidation: Activate this option if you want to perform the k-means algorithm on the partition generated after the truncation (see the description section for more details). The partitions before and after the consolidation are displayed in the Results by object table.

- **Stop conditions**:

- **Iterations**: Enter the maximum number of iterations for the k-means algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 500.

- **Convergence**: Enter the minimum value of evolution for the chosen criterion from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.00001.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlation matrix: Activate this option to display a view of the correlations between the various variables selected.

Proximity matrix: Activate this option to display the proximity matrix.

Node statistics: Activate this option to display the statistics for dendrogram nodes.

Cophenetic distance matrix: Activate this option to display the cophenetic distance matrix.

Cophenetic correlation: Activate this option to display the cophenetic correlation.

Inertia decomposition: Activate this option to display the within-cluster inertia, the between-cluster inertia and the total inertia.

Results by object: Activate this option to display a table giving the cluster each object is assigned to and its distance from the centroids of its cluster (in the initial object order).

- **Correlation with centroids**: Activate this option to display the Pearson correlation between an object and the centroid of its cluster.
    - **Noisy observation**: Activate this option to display a column indicating which observation is noisy. An observation is noisy if the correlation with the centroid of its cluster is smaller than the threshold you choose.
- **Silhouette score**: Activate this option to display the silhouette score of each object.
    - **Mean by cluster**: Activate this option to display a table giving the mean silhouette score for each of the clusters.

Results by cluster: Activate this option to display a table giving the statistics and the objects for each of the clusters.

Centroids: Activate this option to display the table of centroids of the clusters.

Central objects: Activate this option to display the coordinates of the nearest object to the centroid for each cluster.

Charts tab:

Levels bar chart: Activate this option to display the diagram of levels showing the impact of successive clusterings.

Dendrogram: Activate this option to display the dendrogram.

- **Horizontal**: Choose this option to display a horizontal dendrogram.
- **Vertical**: Choose this option to display a vertical dendrogram.
- **Full**: Activate this option to display the full dendrogram (all objects are represented).
- **Truncated**: Activate this option to display the truncated dendrogram (the dendrogram starts at the level of the truncation).
- **Labels**: Activate this option to display object labels (full dendrogram) or clusters (truncated dendrogram) on the dendrogram.
- **Colors**: Activate this option to use colors to represent the different groups on the full dendrogram.

- **Color by group**: Activate this option to represent predefined groups on the dendrogram by coloring the labels according to the selection.

Silhouette score: Activate this option to display a plot showing silhouette score of each object.

Silhouette score (Means): Activate this option to display a plot showing the mean silhouette score of each cluster.

Profile plot: Activate this option to display a plot that allows you to compare the means of the different clusters that have been created.

Results

Summary statistics: This table displays the descriptors of the objects, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation.

Correlation matrix: This table is displayed to give you a view of the correlations between the various variables selected.

Proximity matrix: This table displays the proximities between the objects for the chosen index.

Node statistics: This table shows the data for the successive nodes in the dendrogram. The first node has an index which is the number of objects increased by 1. Hence it is easy to see at any time if an object or group of objects is clustered with another object or group of objects at the level of a new node in the dendrogram.

Levels bar chart: This table displays the statistics for dendrogram nodes.

Cophenetic correlation: This table displays the cophenetic correlation matrix and.

Evolution of the within-cluster variance: If you have selected a number of clusters between two bounds, XLSTAT displays the evolution of the within-cluster variance, which reduces mathematically when the number of clusters increases. If the data is distributed homogeneously, the decrease is linear. If there is actually a group structure, an elbow is observed for the relevant number of clusters.

Evolution of the silhouette score: If you have selected a number of clusters between two bounds, a table with its associated chart shows the evolution of the silhouette score for each k. The optimal number of clusters is the k whose silhouette score is closest to 1.

Evolution of the indices: If you choose **Hartigan (H)**, **Silhouette** or **Calinski and Harabasz** to truncate automatically the dendrogram, a table showing the values of the three indices and the $H(k-1) - H(k)$ value for each number of clusters, is displayed. The value allowing to choose the number of clusters is displayed in bold.

Dendrograms: The full dendrogram displays the progressive clustering of objects. If truncation has been requested, a broken line marks the level the truncation has been carried out. The truncated dendrogram shows the clusters after truncation.

Inertia decomposition: This table shows the within-cluster inertia, the between-cluster inertia and the total inertia.

Cluster centroids: This table shows the cluster centroids for the various descriptors.

Distance between the cluster centroids: This table shows the Euclidean distances between the cluster centroids for the various descriptors.

Central objects: This table shows the coordinates of the nearest object to the centroid for each cluster.

Distance between the central objects: This table shows the Euclidean distances between the cluster central objects for the various descriptors.

Results by cluster: The descriptive statistics for the clusters (number of objects, sum of weights, within-cluster variance, minimum distance to the centroid, maximum distance to the centroid, mean distance to the centroid) are displayed in the first part of the table. The second part shows the objects.

Results by object: This table shows the assignment cluster for each object in the initial object order. * **Cluster**: This column gives the final clustering. It is either the partition obtained after the dendrogram cut or the partition obtained by the k-means algorithm if you activated the consolidation option. * **Cluster (before consolidation)**: This column shows the initial partition generated by the truncated dendrogram. * **Distance to centroid**: This column shows the distance between an object and its cluster centroids.
* **Correlations with centroids**: This column shows the Pearson correlation between an object and its cluster centroids. *NB: it is preferable to interpret the correlation in a standardized space. Indeed, different conclusions can be drawn between the correlation at the centroid and the distance to the centroid, if the data are not at the same scale. Finally, whatever the chosen distance, a centroid is calculated with the Euclidean distance.* * **Noisy observation**: This column indicates which observation is noisy with a bold "Yes" displayed. * **Silhouette scores**: This column shows the silhouette score of each object.

Silhouette scores (Means): This table and its graph are displayed and show the mean silhouette score of each cluster and the silhouette score for the optimal classification (mean of means by cluster).

Profile plot: This chart allows you to compare the means of the different clusters that have been created.

Example

A tutorial on agglomerative hierarchical clustering is available on the XLSTAT Help Center:

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

References

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

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

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 3(1), 1-27.

Hartigan, J. (1975). Clustering algorithms, Wiley, New York.

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, 403-406.

Müllner, D. (2013). fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software*_, 53(1), 1-18.

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.

Gaussian Mixture Models

Use Gaussian mixture models to cluster multidimensional data according to their distribution.

In this section:

Description

Gaussian mixture models allow data to be modeled by a set of Gaussian distributions. Usually, these models are used in a clustering framework and each Gaussian is supposed to correspond to one group.

Mixture model

Denote $x = \{x_1, \ldots, x_n\}$a vector of size n, where $x_i \in R^d$. Assume that each X_i is distributed according to a probability distribution function f:

$$f(x_i; \theta) = \sum_{k=1}^{K} \pi_k h(x_i; \nu_k)$$

where π_k is the mixture proportion of the group k ($\forall k \in \{1, \ldots, K\}, 0 < \pi_k < 1 \text{ and } \sum_{k=1}^{K} \pi_k = 1$) and θ represents the model parameters. The function $h(.; \nu_k)$ is a probability distribution of dimension d with parameter ν_k. For instance, for Gaussian mixture models, h is a Gaussian with mean μ_k and variance Σ_k, hence $\nu_k = (\mu_k, \Sigma_k)$.

Note that, for a mixture distribution, there is a label vector $z = \{z_1, \cdots, z_n\}$with $z_i = \{z_{i1}, \ldots, z_{iK}\}$ defined such that:

$$\begin{cases} z_{ik} = 1 \text{ if } x_i \text{ is assigned to the k-th component} \\ z_{ik} = 0 \text{ otherwise} \end{cases}$$

This vector is often unknown and in a clustering context or density estimation context, the estimation of each z_i is of main interest.

Inference of the model parameters

Due to the latent variables z, the estimation of mixture models parameters cannot be done by directly maximizing the log-likelihood. This optimization requires an iterative algorithm such as the EM (Dempster et al. (1977)) or the SEM, its stochastic version proposed by McLachlan and Peel (2000).

Once the parameters have been estimated, the vector of labels is directly obtained by assigning each x_i to the component providing the highest posterior probability $\hat{\tau}_{ik}$ given by:

$$\hat{\tau}_{ik} = \tau_k(x_i; \hat{\theta}) = \frac{\hat{\pi}_k h(x_i; \hat{\nu}_k)}{\sum_{j=1}^{K} \hat{\pi}_j h(x_i; \hat{\nu}_j)}$$

For a clustering purpose, Celeux and Govaert (1992) proposed the CEM (Classification EM) algorithm which is a k-means-like algorithm and can be viewed as a classifying version of the EM. Contrary to the EM and the SEM, the CEM algorithm maximizes the quantity $\sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log[\pi_k h(x_i; \nu_k)]$ and not the log-likelihood.

Model selection (Choice of the number of components)

The number of components of a mixture model is often unknown. Several criteria such as the BIC (Bayesian Information Criterion, Schwarz (1978)) or the AIC (Akaike Information Criterion, Akaike (1974)) can be used. These criteria are based on a penalization of the observed log-likelihood $L(x; \theta)$. In 2000, Biernacki *et al*. proposed the ICL (Integrated Completed Likelihood) which aims at penalizing the complete log-likelihood $L(x, z; \theta)$. This criterion can be written as a BIC criterion penalized by an entropy term: $-\sum_{i=1}^{n} \sum_{k=1}^{K} \hat{z}_{ik} \log \tau_{ik}$.

For assessing the number of components of a mixture, we can try to find the model which provides well-separated clusters. Proposed by Celeux and Soromengo (1996), the NEC is an entropy-based criterion which measures the overlap of the mixture components:

$$NEC_k = \frac{E_k}{L_k - L_1}$$

where E_k is the entropy of the mixture model with k components and L_k its complete log-likelihood (calculated on the ML estimates). This criterion can also be used as a diagnostic tool. For a given number of components K', if $NEC_{K'} \leq 1$, we can say that there is a clustering structure in the data.

Parsimonious Gaussian mixture models

In the Gaussian mixture models context, the number of parameters can be large and the quantity of data available can be insufficient to achieve a reliable estimate. A classical approach

is to reduce the number of parameters by applying constraints on the variance-covariance matrix Σ_k. Bandfield and Raftery (1993) and Celeux and Govaert (1995) proposed to express the matrix Σ_k in term of its eigenvalue decomposition:

$$\Sigma_k = \lambda_k D_k A_k D_k'$$

where $\lambda_k = |\Sigma_k|^{\frac{1}{d}}$ is the volume of the k-th component, D_k the matrix of eigenvectors and A_k is a diagonal matrix composed by the eigenvalues of $\Sigma_k$organized in decreasing order such that $|A_k| = 1$. These two matrices D_k andA_k allow to control the orientation and the shape of the component.

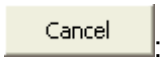
Model	Number of parameters	Model name
$\lambda_k D_k A_k D_k'$	$a + Kb$	VVV
$\lambda DAD'$	$a + b$	EEE
$\lambda D_k A D_k'$	$a + Kb - (K-1)d$	EEV
$\lambda D_k A_k D_k'$	$a + Kb - (K-1)$	EVV
$\lambda_k D_k A D_k'$	$a + Kb - (K-1)(d-1)$	VEV
λB	$a + d$	EEI
λB_k	$a + Kd - K + 1$	EVI
$\lambda_k B_k$	$a + Kd$	VVI
$\lambda_k B$	$a + d + K - 1$	VEI
λI	$a + 1$	EII
$\lambda_k I$	$a + d$	VII
$\lambda_k DAD$	$a + b + K - 1$	VEE
$\lambda DA_k D$	$a + Kb + (K-1)(d-1)$	EVE
$\lambda_k DA_k D$	$a + Kb + (K-1)d$	VVE

Thus, in the multidimensional case, we have 28 different models. In the one- dimensional case, only two models are available (equal variance or not).

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Observations/variables table: Select a table with N objects described by P descriptors. If column headers have been selected, check that the "Column labels" option has been activated.

Row weights: Activate this option if the rows are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column labels" option is activated.

Partial labeling: Activate this option if you want to specify that some rows are constraint to be included in a specific group. If you do not activate this option, all the rows' groups will be considered as unknown. Group identifier must be integers greater than or equal to 1. If a column header has been selected, check that the "Column labels" option is activated.

Data dimension: you can either do a one-dimensional (column by column) or multidimensional analysis.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selections (Observations/variables row weights) contains a label.

Row labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Column labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2…).

Options(1) tab:

Inference algorithm: Select the algorithm used for inference.

- *EM:* Usual EM algorithm proposed by Dempster *et al.* (1977) is used. It is the default algorithm.
- *SEM:* stochastic version of the EM algorithm. A stochastic step is added to the classical EM which aims at assigning the observations to the clusters. This algorithm can lead to empty classes and disturb parameters estimation.
- *CEM:* classifying version of the EM algorithm. A classification step is added to the classical EM which aims at assigning the observations to the clusters. This algorithm can lead to empty classes and disturb parameters estimation.

Selection criteria: Select the criterion to estimate the number of clusters.

- *BIC:* Bayesian Information Criterion. It is the default criterion.
- *AIC:* Akaike Information Criterion. This criterion tends to overestimate the number of components.
- *ICL:* Integrated Complete Likelihood. This criterion searches for the model which provides well-separated clusters. Usually, the selected number of clusters is larger than the number obtained with BIC.
- *NEC:* Normalized Entropy Criterion. The NEC is not defined for a model with one component. This criterion is devoted to choose the number of components rather than the model parameterization.

Initialization: Select the method to initialize the inference algorithm.

- Random: Objects are assigned to classes randomly. The algorithm is run as many times as specified by the number of repetitions until convergence of the algorithm. The best estimate from all repetitions is retained.
- Short runs: Objects are assigned to classes randomly. The algorithm is run as many times as specified by the number of repetitions with a maximum number of 5 iterations. The best estimate from all repetitions is retained to initialize the algorithm.
- K-means: Objects are assigned to classes according to the inference algorithm.

Number of repetitions: Specify the number of repetitions when the initialization method is *Random* or *Short EM*.

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the inference algorithm. The calculations are stopped when the maximum number if iterations has been exceeded. Default value: 500.
- **Convergence**: Enter the minimum value of evolution for the chosen criterion from one iteration to another which, when reached, means that the algorithms is considered to have converged. Default value: 0.00001

Options(2) tab:

Mixture models: Select the model(s) you want to use to fit the data. The best model will be retained according the selection criteria.

Number of classes: Select the minimum and maximum number of classes. The minimum number must be greater than or equal to 1 and the maximum number lower than the number of data points. Default values are 2 and 5.

Equal proportions: Activate this option to constrain mixture proportion to be equal.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Evolution of the criterion: Activate this option to display the evolution table of the chosen criterion.

Posterior probabilities: Activate this option to display the table of posterior probabilities for each cluster.

MAP classification: Activate this option to display the table of classification obtained by the MAP rule.

Charts tab:

Evolution of the criterion: Activate this option to display the evolution chart of the chosen criterion.

MAP classification: Activate this option to display the classification chart obtained by the MAP rule.

Fitted model: Activate this option to display the selected model.

Cumulative density function: Activate this option to display both the empirical and the estimated cdf. This chart is a diagnostic tool. If the two cdf are similar, the mixture model fits well. This chart is only available in the one-dimensional case.

Q-Q plot: Activate this option to display the Q-Q plot of the empirical distribution against the estimated mixture distribution. This chart is a diagnostic tool. If the points in the Q-Q plot approximately lie on the line $y = x$, we can consider the two distributions as similar.

Results

Summary statistics: This table displays for the descriptors of the objects, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation.

Evolution of the criterion: This table displays the values of the criterion for each selected model. A chart is also displayed.

Estimated parameters: Three tables are displayed: the mixture proportions, the means and the variance for each cluster.

Characteristics of the selected model: This table shows some characteristics of the selected model (BIC, AIC, ICL, Log-likelihood, NEC, Entropy, DDL).

Posterior probabilities: The posterior probabilities of belonging to each cluster are displayed in this table.

MAP classification: This table displays the assignment of each observation according to the MAP rule. A chart also displays this classification.

Adjusted model: The fitted model is represented on this chart.

Cumulative density function: This chart allows to compare the empirical cdf to the estimated one.

Q-Q plot: This chart allows to display the quantiles of the empirical distribution against those of the estimated mixture distribution.

Example

A tutorial on Gaussian mixture models is available on XLSTAT Help Center:

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

References

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.

Univariate clustering

Use univariate clustering to optimally cluster objects in k homogeneous classes, based on their description using a single quantitative variable.

In this section:

Description

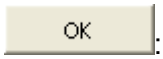
Univariate clustering clusters N one-dimensional observations (described by a single quantitative variable) into k homogeneous classes.

Homogeneity is measured here using the sum of the within-class variances. To maximize the homogeneity of the classes, we therefore try to minimize the sum of the within-class variances.

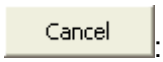
The algorithm used here is very fast and uses the method put forward by W.D. Fisher (1958). This method can be seen as a process of turning a quantitative variable into a discrete ordinal variable. There are many applications, e.g. in mapping applications for creating color scales or in marketing for creating homogeneous segments.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Observations/variables table: Select a table comprising N objects described by P descriptors. If column headers have been selected, check that the "Variable labels" option has been activated.

Row weights: Activate this option if the rows are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column labels" option is activated.

Number of classes: Enter the number of classes to be created by the algorithm.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet in the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selections (Observations/variables table, row labels, row weights, column weights) contains a label.

Row labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Column labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Centroids: Activate this option to display the table of centroids of the classes.

Central objects: Activate this option to display the coordinates of the nearest object to the centroid for each class.

Results by class: Activate this option to display a table giving the statistics and the objects for each of the classes.

Results by object: Activate this option to display a table giving the class each object is assigned to in the initial object order.

Results

Summary statistics: This table displays for the descriptor of the objects, the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation.

Class centroids: This table shows the class centroids for the various descriptors.

Distance between the class centroids: This table shows the Euclidean distances between the class centroids for the various descriptors.

Central objects: This table shows the coordinates of the nearest object to the centroid for each class.

Distance between the central objects: This table shows the Euclidean distances between the class central objects for the various descriptors.

Results by class: The descriptive statistics for the classes (number of objects, sum of weights, within-class variance, minimum distance to the centroid, maximum distance to the centroid, mean distance to the centroid) are displayed in the first part of the table. The second part shows the objects.

Results by object: This table shows the assignment class for each object in the initial object order.

Example

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

References

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

Modeling data

Distribution fitting

Use this tool to fit a distribution to a sample of continuous or discrete quantitative data.

In this section:

Description

Fitting a distribution to a data sample consists, once the type of distribution has been chosen, in estimating the parameters of the distribution so that the sample is the most likely possible (as regards the maximum likelihood) or that at least certain statistics of the sample (mean, variance for example) correspond as closely as possible to those of the distribution.

Distributions

XLSTAT provides the following distributions:

- Arcsine (a): the density function of this distribution (which is a simplified version of the Beta type I distribution) is given by:

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

We have $E(X) = \alpha$ and $V(X) = \alpha(1-\alpha)/2$

- Bernoulli (p): the density function of this distribution is given by:

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

We have $E(X) = p$ and $V(X) = p(1-p)$

The Bernoulli, named after the Swiss mathematician Jacob Bernoulli (1654-1705), allows to describe binary phenomena where only events can occur with respective probabilities of p and

$1-p$.

- Beta (a,b): the density function of this distribution (also called Beta type I) is given by:

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

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

We have $E(X)=\alpha/(\alpha+\beta)$ and $V(X)=\alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2]$

- Beta4 (α,β,c,d): the density function of this distribution is given by:

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

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

We have $E(X)=\frac{c+(c-d)\alpha}{(\alpha+\beta)}$ and $V(X)=\frac{(c-d)^2\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$

Pour the type I beta distribution, X takes values in the $[0,1]$ range. The beta4 distribution is obtained by a variable transformation such that the distribution is on a $[c,d]$ interval where c and d can take any value.

- Binomial (n,p) : the density function of this distribution is given by:

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

We have $E(X)=np$ and $V(X)=np(1-p)$

n is the number of trials, and p the probability of success. The binomial distribution is the distribution of the number of successes for n trials, given that the probability of success is p.

- Negative binomial type I (n,p): the density function of this distribution is given by:

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

We have $E(X)=n(1-p)/p$ and $V(X)=n(1-p)/p^2$

n is the number of successes, and p the probability of success. The negative binomial type I distribution is the distribution of the number x of unsuccessful trials necessary before obtaining n successes.

- Negative binomial type II (k,p): the density function of this distribution is given by:

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

We have $E(X)=kp$ and $V(X)=kp(p+1)$

The negative binomial type II distribution is used to represent discrete and highly heterogeneous phenomena. As k tends to infinity, the negative binomial type II distribution tends towards a Poisson distribution with $\lambda=kp$.

- $Khi^2(df)$: the density function of this distribution is given by:

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

We have $E(X)=df$ and $V(X)=2df$

The Chi-square distribution corresponds to the distribution of the sum of df squared standard normal distributions. It is often used for testing hypotheses.

- Erlang (k,λ): the density function of this distribution is given by:

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

We have $E(X)=k/\lambda$ and $V(X)=k/\lambda^2$

k is the shape parameter and λ is the rate parameter.

This distribution, developed by the Danish scientist A. K. Erlang (1878-1929) when studying the telephone traffic, is more generally used in the study of queuing problems.

Note: When $k=1$, this distribution is equivalent to the exponential distribution. The Gamma distribution with two parameters is a generalization of the Erlang distribution to the case where k is a real and not an integer (for the Gamma distribution the scale parameter $\beta=1/\lambda$ is used).

- Exponential(λ): the density function of this distribution is given by:

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

We have $E(X)=1/\lambda$ and $V(X)=1/\lambda^2$

The exponential distribution is often used for studying lifetime in quality control.

- Fisher (df_1,df_2): the density function of this distribution is given by:

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

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

We have $E(X) = df_2/(df_2 - 2)$ if $df_2 > 2$, and $V(X) = \frac{2df_2^2(df_1 + df_2 - 2)}{df_1(df_2-2)^2(df_2-4)}$ if $df_2 > 4$

Fisher's distribution, from the name of the biologist, geneticist and statistician Ronald Aylmer Fisher (1890-1962), corresponds to the ratio of two Chi-square distributions. It is often used for testing hypotheses.

- Fisher-Tippett (β, μ): the density function of this distribution is given by:

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

We have $E(X) = \mu + \beta\gamma$ and $V(X) = (\pi\beta)^2/6$ where γ is the Euler-Mascheroni constant.

The Fisher-Tippett distribution, also called the Log-Weibull or extreme value distribution, is used in the study of extreme phenomena. The Gumbel distribution is a special case of the Fisher-Tippett distribution where $\beta = 1$ and $\mu = 0$.

- Gamma (k, β, μ): the density of this distribution is given by:

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

We have $E(X) = \mu + k\beta$ and $V(X) = k\beta^2$

k is the shape parameter of the distribution and β the scale parameter.

- GEV (β, k, μ): the density function of this distribution is given by:

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

We have $E(X) = \mu + \frac{\beta}{k}\Gamma(1+k)$ and $V(X) = \left(\frac{\beta}{k}\right)^2 \left(\Gamma(1+2k) - \Gamma^2(1+k)\right)$

The GEV (Generalized Extreme Values) distribution is much used in hydrology for modeling flood phenomena. k lies typically between -0.6 and 0.6.

- Gumbel: the density function of this distribution is given by:

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

We have $E(X) = \gamma$ and $V(X) = \pi^2/6$ where γ is the Euler-Mascheroni constant (0.5772156649…).

The Gumbel distribution, named after Emil Julius Gumbel (1891-1966), is a special case of the Fisher-Tippett distribution with $\beta = 1$ and $\mu = 0$. It is used in the study of extreme phenomena such as precipitations, flooding and earthquakes.

- Logistic (μ, s): the density function of this distribution is given by:

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

We have $E(X) = \mu$ and $V(X) = (\pi s)^2/3$

- Lognormal (μ, σ): the density function of this distribution is given by:

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

We have $E(X) = exp(\mu + \sigma^2/2)$ and $V(X) = \left(exp(\sigma^2) - 1\right) exp(2\mu + \sigma^2)$

- Lognormal2 (m, s): the density function of this distribution is given by:

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

We have:

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

And:

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

This distribution is just a reparametrization of the Lognormal distribution.

- Normal (μ, σ) : the density function of this distribution is given by:

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

We have $E(X) = \mu$ and $V(X) = \sigma^2$

- Standard normal: the density function of this distribution is given by:

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

We have $E(X) = 0$ and $V(X) = 1$

This distribution is a special case of the normal distribution with $\mu = 0$ and $\sigma = 1$

- Pareto (a, b): the density function of this distribution is given by:

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

We have $E(X) = ab/(a-1)$ with $V(X) = \frac{ab^2}{((a-1)^2(a-2))}$

The Pareto distribution, named after the Italian economist Vilfredo Pareto (1848-1923), is also known as the Bradford distribution. This distribution was initially used to represent the distribution of wealth in society, with Pareto's principle that 80% of the wealth was owned by 20% of the population.

- PERT (a, m, b): the density function of this distribution is given by:

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

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

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

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

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

The PERT distribution is a special case of the beta4 distribution. It is defined by its definition interval [a, b] and m the most likely value (the mode). PERT is an acronym for *Program Evaluation and Review Technique*, a project management and planning methodology. The PERT methodology and distribution were developed during the project held by the US Navy and Lockheed between 1956 and 1960 to develop the Polaris missiles launched from submarines. The PERT distribution is useful to model the time that is likely to be spent by a team to finish a project. The simpler triangular distribution is similar to the PERT distribution in that it is also defined by an interval and a most likely value.

- Poisson (λ): the density function of this distribution is given by:

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

We have $E(X) = \lambda$ with $V(X) = \lambda$

Poisson's distribution, discovered by the mathematician and astronomer Siméon- Denis Poisson (1781-1840), pupil of Laplace, Lagrange and Legendre, is often used to study queuing

phenomena.

- Student (df) : the density function of this distribution is given by:

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

We have $E(X) = 0$ if $df > 1$ with $V(X) = df/(df-2)$ if $df > 2$

The English chemist and statistician William Sealy Gosset (1876-1937), used the nickname Student to publish his work, in order to preserve his anonymity (the Guinness brewery forbade its employees to publish following the publication of confidential information by another researcher). The Student's t distribution is the distribution of the mean of df variables standard normal variables. When $df = 1$, Student's distribution is a Cauchy distribution with the particularity of having neither expectation nor variance.

- Trapezoidal (a, b, c, d): the density function of this distribution is given by:

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

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

This distribution is useful to represent a phenomenon for which we know that it can take values between two extreme values (a and d), but that it is more likely to take values between two values (b and c) within that interval.

- Triangular (a, m, b): the density function of this distribution is given by:

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

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

- TriangularQ (q_1, m, q_2, p_1, p_2): the density function of this distribution is a reparametrization of the Triangular distribution. A first step requires estimating the a and b parameters of the triangular distribution, from the q_1 and q_2 quantiles to which percentages p_1 and p_2 correspond. Once this is done, the distribution functions can be computed using the triangular distribution functions.

- Uniform (a, b): the density function of this distribution is given by:

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

We have $E(X) = (a + b)/2$ with $V(X) = (b - a)^2/12$

The uniform (0,1) distribution is much used for simulations. As the cumulative distribution function of all the distributions is between 0 and 1, a sample taken in a Uniform (0,1) distribution is used to obtain random samples in all the distributions for which the inverse can be calculated.

- Uniform discrete (a, b): the density function of this distribution is given by:

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

We have $E(X) = (a + b)/2$ with $V(X) = [(b - a + 1)^2 - 1]/12$

The uniform discrete distribution corresponds to the case where the uniform distribution is restricted to integers.

- Weibull (β): the density function of this distribution is given by:

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

We have $E(X) = \Gamma(\frac{1}{\beta} + 1)$ with $V(X) = \Gamma(\frac{2}{\beta} + 1) - \Gamma^2(\frac{1}{\beta} + 1)$

β is the shape parameter for the Weibull distribution.

- Weibull (β, γ): the density function of this distribution is given by:

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

We have $E(X) = \gamma\Gamma\left(\frac{1}{\beta}+1\right)$ with $V(X) = \gamma^2\left[\Gamma\left(\frac{2}{\beta}+1\right) - \Gamma^2\left(\frac{1}{\beta}+1\right)\right]$

β is the shape parameter of the distribution and γ the scale parameter. When $\beta = 1$, the Weibull distribution is an exponential distribution with parameter $1/\gamma$.

- Weibull (β, γ, μ): the density function of this distribution is given by:

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

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

The Weibull distribution, named after the Swede Ernst Hjalmar Waloddi Weibull (1887-1979), is much used in quality control and survival analysis. β is the shape parameter of the distribution and γ the scale parameter. When $\beta = 1$ and $\mu = 0$, the Weibull distribution is an exponential distribution with parameter $1/\gamma$.

Fitting method

XLSTAT offers two fitting methods:

Moments: this simple method uses the definition of the moments of the distribution as a function of the parameters to determine the latter. For most distributions, the use of the mean and the variance is sufficient. However, for certain distributions, the mean suffices (for example Poisson's distribution), or, if not, the asymmetry coefficient is also required (Weibull's distribution for example).

Likelihood: the parameters of the distribution are estimated by maximizing the likelihood of the sample. This method, more complex, has the advantage of rigor for all distributions and enables approximate standard deviations to be obtained for parameter estimators. The maximum likelihood method is offered for the negative binomial type II distribution, Fisher-Tippett distribution, GEV distribution and Weibull distribution.

For certain distributions, the moments method gives exactly the same result as the maximum likelihood method. This is particularly true for the normal distribution.

Goodness of fit test

Once the parameters for the chosen distribution have been estimated, the hypothesis must be tested in order to check if the phenomenon observed through the sample follows the distribution in question. XLSTAT offers two goodness of fit tests.

The **Chi-square goodness of fit test** is a parametric test using the distance (as regards Chi-square) between the histogram of the theoretical distribution (determined by the estimated parameters) and the histogram of the empirical distribution of the sample. The histograms are calculated using k intervals chosen by the user. It is shown that the statistic calculated asymptotically follows a Chi-square distribution with (n-k) degrees of freedom where n is the number of observations in the sample. This test is better for discrete distributions and it is recommended to check that the expected frequency in each class is not less than 5.

It may happen that the Chi-square test leads to a bad fit of the distribution to the data with one class contributing much more to the Chi-square than the others. In this case, the union of the class in question with a neighboring class is used to check if the conclusion is due only to the class in question or it is actually the fit which is incorrect.

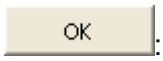
The **Kolmogorov-Smirnov goodness of fit test** is an exact non-parametric test based on the maximum distance between a theoretical distribution function (entirely determined by the known values of its parameters) and the empirical distribution function of the sample. This test can only be used for continuous distributions.

When a parameter estimation precedes the goodness of fit test, the Kolmogorov- Smirnov test is not correct as the parameters are estimated by trying to bring the theoretical distribution as close as possible to the data. If it confirms the goodness of fit hypothesis, the Kolmogorov-Smirnov test risks being optimistic.

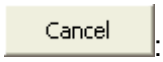
For the case where the distribution used is the normal distribution, Lilliefors and Stephens (see normality tests) have put forward a modified Kolmogorov-Smirnov test which allows parameters to be estimated on the sample tested.

Dialog box

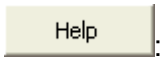
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data: Select the data for which the goodness of fit test is to be calculated. You can select several columns (columns mode) or rows (rows mode) if you want to carry out tests on several samples at the same time.

Distribution: Choose the probability distribution to be used for the fit and/or goodness of fit tests. See the description section for more information on the distributions offered. The **automatic** option allows to let XLSTAT identify the best fitting distribution (determined using a Kolmogorov-Smirnov test).

Parameters: You can choose to **enter** the parameters for the distribution, or **estimate** them. If you choose to enter the parameters, you must enter their values.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Sample labels: Activate this option if the sample labels are on the first row (columns mode) or in the first column (rows mode) of the selected data.

Weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

- **Standardize weights**: If you activate this option, the weights are standardized such that their sum equals the number of observations.

Options tab:

Tests: Choose the type of goodness of fit test (see the description section for more details on the tests).

- **Kolmogorov-Smirnov**: Activate this option to perform a Kolmogorov-Smirnov test.
- **Chi-square**: Activate this option to perform the Chi-square test.
- **Significance level (%)**: Enter the significance level for the above tests.

Estimation method: Choose the method of estimating the parameters of the chosen distribution (see the description section for more details on estimation methods).

- **Moments**: Activate this option to use the moments method.

- **Maximum likelihood**: Activate this option to use the maximum likelihood method. You can then change the convergence limit value which when reached means the algorithm is considered to have converged. Default value: 0.00001.

Intervals: For a Chi-square distribution, or if you want to compare the density of the distribution chosen with the sample histogram, you must choose one of the following options:

- **Number**: Choose this option to enter the number of intervals to create.
- **Width**: Choose this option to define a fixed width for the intervals.
- **User defined**: Select a column containing in increasing order the lower bound of the first interval, and the upper bound of all the intervals.
- **Minimum**: Activate this option to enter the value of the lower value of the first interval. This value must be lower or equal to the minimum of the series.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the samples selected.

Charts tab:

Histograms: Activate this option to display the histograms of the samples. For a theoretical distribution, the density function is displayed.

- **Bars**: Choose this option to display the histograms with a bar for each interval.
- **Continuous lines**: Choose this option to display the histograms with a continuous line.

Cumulative histograms: Activate this option to display the cumulative histograms of the samples. For a theoretical distribution, the distribution function is displayed.

- **Based on the histogram**: Choose this option to display cumulative histograms based on the same interval definition as the histograms.
- **Empirical cumulative distribution**: Choose this option to display cumulative histograms which actually correspond to the empirical cumulative distribution of the sample.

Results

Summary statistics: This table displays for the selected samples, the number of observations, the number of missing values, the number of non- missing values, the mean and the standard deviation.

Estimated parameters: This table displays the parameters for the distribution.

Statistics estimated on the input data and computed using the estimated parameters of the distribution: This table is used to compare the mean, variance, skewness and kurtosis coefficients calculated from the sample with those calculated from the values of the distribution parameters.

Kolmogorov-Smirnov test: The results of the Kolmogorov-Smirnov test are displayed if the corresponding option has been activated.

Chi-square test: The results of the Chi-square test are displayed if the corresponding option has been activated.

Comparison between the observed and theoretical frequencies: This table is displayed if a Chi-square test was requested.

Descriptive statistics for the intervals: This table is displayed if histograms have been requested. It shows for each interval the frequencies, the relative frequencies, together with the densities for the samples and distribution chosen.

Example

A tutorial on distribution fitting is available on the XLSTAT Help Center:

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

References

Abramowitz M. & I.A. Stegun (1972). Handbook of Mathematical Functions. Dover Publications, New York, 925-964.

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.

Linear regression

Use this tool to create a simple or multiple linear regression model for explanation or prediction.

In this section:

Description

Linear regression is, without doubt, the most frequently used statistical method. A distinction is usually made between simple regression (with only one explanatory variable) and multiple regression (several explanatory variables) although the overall concept and calculation methods are identical.

The principle of linear regression is to model a quantitative dependent variable Y through a linear combination of p quantitative explanatory variables, $X_1, X_2, \ldots, X_p$. The determinist model (not taking randomness into account) is written for observation i as follows:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \qquad (1)$$

where y_i is the value observed for the dependent variable for observation i, x_{ij} is the value taken by variable j for observation i, and ϵ_i is the error of the model.

The statistical framework and the hypotheses which accompany it are not required for fitting this model. Furthermore, minimization using the least squares method (the sum of squared errors ϵ_i^2 is minimized) provides an exact analytical solution. However, to be able to test the hypotheses and measure the explanatory power of the various explanatory variables in the model, a statistical framework is necessary.

The linear regression hypotheses are as follows: the errors e_i follow the same normal distribution $N(0, s)$ and are independent.

The way the model with this hypothesis added is written means that, within the framework of the linear regression model, the y_i are the expression of random variables with mean μ_i and variance s^2, where

$$\mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \quad (2)$$

The estimator of the β coefficients and of their covariance matrix are given by:

$$\hat{\beta} = (X^t X)^{-1} X^t Y \quad (3)$$

and

$$Var(\hat{\beta}) = \hat{\sigma}^2 (X^t X)^{-1} \quad (4)$$

To use the various tests proposed in the results of linear regression, it is recommended to check retrospectively that the underlying hypotheses have been correctly verified. The normality of the residuals can be checked by analyzing certain charts or by using a normality test. The independence of the residuals can be checked by analyzing certain charts or by using the Durbin-Watson test.

Correcting for Heteroscedasticity and Autocorrelation

Homoscedasticity and independence of the error terms are key hypotheses in linear regression where it is assumed that the variances of the error terms are independent and identically distributed and normally distributed. When these assumptions are not possible to keep (a Durbin Watson or White test available in the Time series menu allow to challenge these hypotheses), a consequence is that the covariance matrix cannot be estimated using the classical formula, and the variance of the parameters corresponding to the β coefficients of the linear model can be wrong and their confidence intervals as well. A predictor could be said to be significant (or respectively not) while being the opposite. XLSTAT allows to correct for heteroscedasticity and autocorrelation that can arise, especially in time series.

For what concerns heteroscedasticity, White (1980) followed by several authors has explored ways to correct the classical estimate of the covariances using residuals and centered leverages obtained from the linear regression computations (see MacKinnon (1985) and Zeileis (2006) for a review). When the assumptions of classical linear regression do not hold, while the estimators of the coefficients are unchanged, the simplified writing of the covariance matrix (see equation 4) of the beta parameters is not possible, and we must revert to the general expression:

$$Var(\beta) = (X^t X)^{-1} (X^t \Omega X) (X^t X)^{-1} \quad (5)$$

Equation (5) is equivalent to equation (4) when

$$\Omega = \hat{\sigma}^2 I \quad (6)$$

Let ω_i be the diagonal elements of W. The different heteroscedasticity coefficients (HC) estimators for the ω_i are given by:

$$HC0 : \omega_i = {\hat{e_i}}^2$$
$$HC1 : \omega_i = {\hat{e_i}}^2 \frac{n}{(n-p-1)}$$
$$HC2 : \omega_i = \frac{{\hat{e_i}}^2}{(1-h_i)}$$
$$HC3 : \omega_i = \frac{{\hat{e_i}}^2}{(1-h_i)^2}$$
$$HC4 : \omega_i = \frac{{\hat{e_i}}^2}{(1-h_i)^{\delta_i}} \text{ with } \delta_i = min(4, h_i/\bar{h})$$

where $\hat{e_i}$ is the residual, and h_i the centered leverage for the ith observation, and p is the number of predictors.

Newey and West (1987) suggested an estimator that corrects for both autocorrelation and heteroscedasticity, but the lag must be known from the user (the descriptive analysis of time series or ARIMA functions of XLSTAT can be used for that). For lag=0 (no autocorrelation) we have:

$$X^t \Omega X = X^t \Omega_0 X = \frac{n}{n-p-1} \sum_{i=1}^{n} \hat{e}_i^2 x_i^t x_i$$

where x_i is the vector of the predictors (including a 1 for the intercept of the model) for the ith observation. For lag m (m>0), we have:

$$X^t \Omega X = X^t \Omega_0 X + \frac{n}{n-p-1} \sum_{l=1}^{m} \sum_{t=l+1}^{n} \hat{e}_t^2 \hat{e}_{t-l}^2 (x_t^t x_{t-l} - x_{t-l}^t x_t)$$

The unadjusted version of the Newey West estimator corresponds to the same approach without the $n/(n-p-1)$ adjustment factor.

The **Clusters** option makes it possible to correct the problem of heteroscedasticity in the case where the variances are considered to be equal only within given specific clusters. When this option is selected, you must then select the data indicating to which cluster each observation belongs.

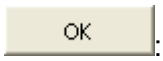
$$X^t \Omega X = \frac{n-1}{n-p-1} \frac{K}{K-1} \sum_{g=1}^{K} X_g^t \hat{\mathbf{e}} \hat{\mathbf{e}}^{\mathbf{t}} X_g$$

where K is the number of clusters and X_g is the subset of observations belonging to the g^{th} cluster.

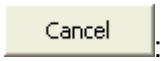
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various

elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Y / Dependent variables:

Quantitative: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

X / Explanatory variables:

Quantitative: Select the quantitative explanatory variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Activate this option to perform an ANCOVA analysis. Then select the qualitative explanatory variables (the factors) in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

Regression weights: Activate this option if you want to carry out a weighted least squares regression. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Groups: Activate this option if you want to group the data. Then select the data that correspond to the group to which each observation belongs.

Regression by group: Activate this option if you want the calculations to be performed on each group separately. If this option is not checked, the groups will appear in different colors on the graphs, but the regression will be done on the entire data.

Options tab:

Model sub-tab:

Fixed constant: Activate this option to fix the constant of the regression model to a value you then enter (0 by default).

Tolerance: Activate this option to prevent the model from taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

Model selection: Activate this option if you want to use one of the four selection methods provided:

- **Best model**: This method lets you choose the best model from amongst all the models which can handle a number of variables varying from "Min variables" to "Max Variables". Furthermore, the user can choose several "criteria" to determine the best model.
- **Criterion**: Choose the criterion from the following list: Adjusted R², Mean Square of Errors (MSE), Mallows Cp, Akaike's AIC, Schwarz's SBC, Amemiya's PC.
- **Min variables**: Enter the minimum number of variables to be used in the model.
- **Max variables**: Enter the maximum number of variables to be used in the model.

Note: this method can cause long calculation times as the total number of models explored is the sum of the (Cn, k)s for k varying from "Min variables" to "Max variables", where (Cn, k) is equal to $\frac{n!}{(n-k)!k!}$. It is there recommended that the value of "Max variables" be increased gradually.

- **Stepwise**: The selection process starts by adding the variable with the largest contribution to the model (the criterion used is Student's t statistic). If a second variable is such that the probability associated with its$t$is less than the "**Probability for entry** ", it is added to the model. The same for a third variable. After the third variable is added, the impact of removing each variable present in the model after it has been added is evaluated (still using the t statistic). If the probability is greater than the "**Probability of removal** ", the variable is removed. The procedure continues until no more variables can be added or removed.
- **Forward**: The procedure is the same as for stepwise selection except that variables are only added and never removed.
- **Backward**: The procedure starts by simultaneously adding all variables. The variables are then removed from the model following the procedure used for stepwise selection.

Covariances sub-tab:

In this tab you can choose to apply corrections for heteroscedasticity and autocorrelation. See the *description* section for computational details.

Validation tab:

Validation: Activate this option if you want to use a sub-sample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.

- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

Quantitative: Activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

Qualitative: Activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

Observations labels: activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

- **Check for each Y separately**: Choose this option to remove the observations with missing data in the select Y (dependent) variables, only if the Y of interest has a missing data.
- **Across all Ys**: Choose this option to remove the observations with missing data in the Y (dependent) variables, even if the Y of interest has no missing data.

Ignore missing data: Activate this option to ignore missing data. If missing data are present for the dependent variable, the corresponding observations will be predicted. If missing data are present for the explanatory variable(s) the corresponding observations are used to estimate the correlation matrix with pairwise deletion.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

General sub-tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the correlation matrix for quantitative variables (dependent or explanatory).

Effect size measures : activate this option to display the effect size measures. As part of the linear regression, the following measures are displayed: * R^2 * $f^2 = \frac{R^2}{1-R^2}$

Multicolinearity statistics: Activate this option to display the multicolinearity statistics for all explanatory variables.

Analysis of variance: Activate this option to display the analysis of variance table.

Type I/III SS: Activate this option to display the Type I and Type III sum of squares tables.

Press: Activate this option to calculate and display the Press (predicted residual error sum of squares) statistic.

Interpretation: Activate this option to display an automatic interpretation of the regression results.

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

- **X**: Activate this option to display the explanatory variables in the predictions and residuals table.
- **Confidence intervals**: Activate this option to calculate and display the confidence intervals on the predictions.
- **Adjusted predictions**: Activate this option to calculate and display adjusted predictions in the table of predictions and residuals.

- **Influence diagnostics**: Activate this option to calculate and display the table that contains the influence statistics for each observation.

Contrasts sub-tab:

Compute contrasts: Activate this option to compute contrasts, then select the contrasts table, where there must be one column per contrast and one row for each coefficient of the model.

Test assumptions sub-tab:

This option is only available if the **Prediction and residuals** have been requested in the Outputs/General tab.

Normality test: Activate this option to run a Shapiro-Wilk test on the residuals.

Charts tab:

Regression charts: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.
- **Predictions and residuals**: Activate this option to display the following charts.

(1) Line of regression: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

(2) Explanatory variable versus standardized residuals: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

(3) Dependent variable versus standardized residuals.

(4) Predictions for the dependent variable versus the dependent variable.

(5) Bar chart of standardized residuals.

Confidence intervals: Activate this option to have confidence intervals displayed on charts (1) and (4).

Results

Summary statistics: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the dependent variables (in green) and the quantitative explanatory variables. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

Correlation matrix: This table is displayed to give you a view of the correlations between the various variables selected.

Summary of the variables selection: Where a selection method has been chosen, XLSTAT displays the selection summary. For a stepwise selection, the statistics corresponding to the

different steps are displayed. Where the best model for a number of variables varying from p to q has been selected, the best model for each number or variables is displayed with the corresponding statistics and the best model for the criterion chosen is displayed in bold.

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.
- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).
- **R²**: The determination coefficient for the model. This coefficient, which value is between 0 and 1, is defined by:

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ with } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

The R² is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer R² is to 1, the better is the model. The problem with the R² is that it does not take into account the number of variables used to fit the model.

- **Adjusted R²**: The adjusted determination coefficient for the model. The adjusted R² can be negative if the R² is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}$$

The adjusted R² is a correction to the R² which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W - p^*}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.
- **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows:

$$MAPE = \frac{100}{W}\sum_{i=1}^{n} w_i \left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

- **DW**: The Durbin-Watson statistic is defined by:

$$DW = \frac{\sum_{i=2}^{n}[(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2}$$

This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- **Cp**: Mallows Cp coefficient is defined by:

$$Cp = \frac{SSE}{\hat{\sigma}} + 2p^* - W$$

where SSE is the sum of the squares of the errors for the model with p explanatory variables and $\hat{\sigma}$ is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to p*, the less the model is biased.

- **AIC**: Akaike's Information Criterion is defined by:

$$AIC = W \ln(\frac{SSE}{W}) + 2p^*$$

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz's Bayesian Criterion is defined by:

$$SBC = W \ln(\frac{SSE}{W}) + \ln(W)p^*$$

This criterion, proposed by Schwarz (1978) is similar to the AIC, and the aim is to minimize it.

- **PC**: Amemiya's Prediction Criterion is defined by:

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}$$

This criterion, proposed by Amemiya (1980) is used, like the adjusted R² to take account of the parsimony of the model.

- **Press**: The Press (predicted residual error sum of squares) statistic is only displayed if the corresponding option has been activated in the dialog box. It is defined by:

$$Press = \sum_{i=1}^{n} w_i(y_i - \hat{y}_{i(-i)})^2$$

where $\hat{y}_{i(-i)}$is the prediction for observation i when the latter is not used for estimating parameters. We then get:

$$\text{Press } RMSE = \sqrt{\frac{Press}{W - p^*}}$$

The Press RMSE can then be compared to the RMSE. A large difference between the two shows that the model is sensitive to the presence or absence of certain observations in the model.

- **Q²**: This statistic, also known as the cross-validated R², is only displayed if the Press option has been activated in the dialog box. It is defined by:

$$Q^2 = 1 - \frac{\text{Press}}{\sum_{i=1}^{x}(y_i - \overline{y})^2}$$

This gives the proportion of the total variance that is explained by the explanatory variables when the predictions are computed when the corresponding observation is not in the model. A large difference between the Q² and the R² shows that the model is sensitive to the presence or absence of certain observations in the model.

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

If the Type I/III SS (SS: Sum of Squares) is activated, the corresponding tables are displayed.

The table of **Type I SS** (sum of squares) values is used to visualize the influence that progressively adding explanatory variables has on the fitting of the model, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. The sums of squares in the Type I table always add up to the model SS. Note: the order in which the variables are selected in the model influences the values obtained.

The table of **Type II SS** (sum of squares) values is used to visualize the influence that removing an explanatory variable has on the fitting of the model (if included in the model, the interactions that depend on the variable are removed as well), all other variables being retained, expect those were the effect is present (interactions), as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. Type II and Type III are identical if there are no interactions or if the design is balanced.

The table of **Type III SS** (sum of squares) values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained (including interactions including the variable in question), expect those were the effect is present (interactions), as regards the sum of the squares of the errors (SSE), the mean squared error

(MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. Type II and Type III are identical if there are no interactions or if the design is balanced.

The **parameters of the model** table displays the estimate of the parameters, the corresponding standard error, the Student's t, the corresponding probability, as well as the confidence interval.

The **equation of the model** is then displayed to make it easier to read or re-use the model.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of normalized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the observed value of the dependent variable, the model's prediction, the residuals, the confidence intervals together with the adjusted prediction if the corresponding options have been activated in the dialog box. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the uncertainty being larger. If validation data have been selected, they are displayed at the end of the table.

The **charts** which follow show the results mentioned above. If there is only one explanatory variable in the model, the first chart displayed shows the observed values, the regression line and both types of confidence interval around the predictions. The second chart shows the standardized residuals as a function of the explanatory variable. In principle, the residuals should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a problem with the model.

The **three charts** displayed next show respectively the evolution of the standardized residuals as a function of the dependent variable, the distance between the predictions and the observations (for an ideal model, the points would all be on the bisector), and the standardized residuals on a bar chart. The last chart quickly shows if an abnormal number of values are outside the interval]-2, 2[given that the latter, assuming that the sample is normally distributed, should contain about 95% of the data.

The table of **influence diagnostics** displays for each observation, its weight, the corresponding residual, the standardized residual (divided by the $RMSE$), the studentized residual, the deleted residual, the studentized deleted residual, the centered leverage, the Mahalanobis distance, the Cook's D, the CovRatio, the DFFit, the standardized DFFit, the DFBetas (one per model coefficient) and the standardized DFBetas.

Four charts are then displayed to make possible an easy identification of the observations which influence on the predictions or on the coefficients make necessary a special investigation.

If the **tests on normality** has been requested, the corresponding results are then displayed.

If you have selected the data to be used for calculating **predictions on new observations**, the corresponding table is displayed next.

Example

A tutorial on simple linear regression is available on the XLSTAT Help Center:

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

A tutorial on multiple linear regression is available on XLSTAT Help Center:

http://www.xlstat.com/demo-reg2.htm

References

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.

Jobson J. D. (1999). Applied Multivariate Data Analysis: Volume 1: Regression and Experimental Design. Springer Verlag, New York.

MacKinnon J. G. and White H. (1985). Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties. *Journal of Econometrics*, **29**, 305-325.

Mallows C.L. (1973). Some comments on Cp. *Technometrics*, **15**, 661-675.

Newey W.K.; West K.D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, **55 (3)**: 703-708.

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.

Welch B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, **38**, 330-336.

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

Use this model to carry out ANOVA (ANalysis Of VAriance) of one or more balanced or unbalanced factors. The advanced options enable you to choose the constraints on the model and to take account of interactions between the factors. Multiple comparison tests can be calculated.

In this section:

Description

Analysis of variance (ANOVA) uses the same conceptual framework as linear regression. The main difference comes from the nature of the explanatory variables: instead of quantitative, here they are qualitative. In ANOVA, explanatory variables are often called factors.

If p is the number of factors, the ANOVA model is written as follows:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_{k(i,j),j} + \epsilon_i \quad (1)$$

where y_i is the value observed for the dependent variable for observation i, k(i,j) is the index of the category of factor j for observation i, and ϵ_i is the error of the model.

The hypotheses used in ANOVA are identical to those used in linear regression: the errors ϵ_i follow the same normal distribution $N(0, s)$ and are independent.

The way the model with this hypothesis added is written means that, within the framework of the linear regression model, the y_i are the expression of random variables with mean μ_i and variance s^2, where

$$\mu_i = \beta_0 + \sum_{j=1}^{p} \beta_{k(i,j),j} \quad (2)$$

The estimator of the β coefficients and of their covariance matrix are given by:

$$\hat{\beta} = (X^t X)^{-1} X^t Y \quad (3)$$

and

$$Var(\hat{\beta}) = \hat{\sigma}^2 (X^t X)^{-1} \quad (4)$$

To use the various tests proposed in the results ANOVA, it is recommended to check retrospectively that the underlying hypotheses have been correctly verified. The normality of the residuals can be checked by analyzing certain charts, by using a normality test, or by running a Levene's test to test if the variances for the different groups are homogeneous. The independence of the residuals can be checked by analyzing certain charts or by using the Durbin Watson test.

Data selection

Typically, in order to run an ANOVA in XLSTAT, you need to enter each variable in a single column.

XLSTAT allows you to select the data in two different ways when having up to three factors (explanatory variables):

- Select a single column of values for each variable (dependent and factors).
- Select a table where rows categorize the data according to one factor, and columns categorize them according to the other factors (1 or 2 variables). In this case, it is impossible to estimate the missing data. They will therefore necessarily be deleted.

Correcting for Heteroscedasticity and Autocorrelation

Homoscedasticity and independence of the error terms are key hypotheses in linear regression and ANOVA where it is assumed that the variances of the error terms are independent and identically distributed and normally distributed. When these assumptions are not possible to keep (a Durbin Watson or White test available in the Time series menu allow to challenge these hypotheses), a consequence is that the covariance matrix cannot be estimated using the classical formula, and the variance of the parameters corresponding to the coefficients of the linear model can be wrong and their confidence intervals as well. A predictor could be said to be significant (or respectively not) while being the opposite. XLSTAT allows to correct for heteroscedasticity and autocorrelation that can arise, especially in time series.

For what concerns heteroscedasticity, White (1980) followed by several authors has explored ways to correct the classical estimate of the covariances using residuals and centered leverages obtained from the linear regression computations (see MacKinnon (1985) and Zeileis (2006) for a review). When the assumptions of classical linear regression do not hold, while the estimators of the coefficients are unchanged, the simplified writing of the covariance matrix (see equation 4) of the β parameters is not possible, and we must revert to the general expression:

$$Var(\beta) = (X^t X)^{-1} (X^t \Omega X) (X^t X)^{-1} \quad (5)$$

Equation (5) is equivalent to equation (4) when

$$\Omega = \hat{\sigma}^2 I \quad (6)$$

Let ω_i be the diagonal elements of W. The different heteroscedasticity coefficients (HC) estimators for the ω_i are given by:

$$HC0: \quad \omega_i = \hat{e_i}^2$$
$$HC1: \quad \omega_i = \hat{e_i}^2 \frac{n}{(n-p-1)}$$
$$HC2: \quad \omega_i = \frac{\hat{e_i}^2}{(1-h_i)}$$
$$HC3: \quad \omega_i = \frac{\hat{e_i}^2}{(1-h_i)^2}$$
$$HC4: \quad \omega_i = \frac{\hat{e_i}^2}{(1-h_i)^{\delta_i}} \text{ with } \delta_i = min(4, h_i/\bar{h})$$

where $\hat{e_i}$ is the residual, and h_i the centered leverage for the ith observation, and p is the number of predictors.

Newey and West (1987) suggested an estimator that corrects for both autocorrelation and heteroscedasticity, but the lag must be known from the user (the descriptive analysis of time series or ARIMA functions of XLSTAT can be used for that). For lag=0 (no autocorrelation) we have:

$$X^t \Omega X = X^t \Omega_0 X = \frac{n}{n-p-1} \sum_{i=1}^{n} \hat{e}_i^2 x_i^t x_i$$

where x_i is the vector of the predictors (including a 1 for the intercept of the model) for the ith observation. For lag m (m>0), we have:

$$X^t \Omega X = X^t \Omega_0 X + \frac{n}{n-p-1} \sum_{l=1}^{m} \sum_{t=l+1}^{n} \hat{e}_t^2 \hat{e}_{t-l}^2 (x_t^t x_{t-l} - x_{t-l}^t x_t)$$

The unadjusted version of the Newey West estimator corresponds to the same without the $n/(n-p-1)$ adjustment factor.

The **Clusters** option makes it possible to correct the problem of heteroscedasticity in the case where the variances are considered to be equal only within given specific clusters. When this option is selected, you must then select the data indicating to which cluster each observation belongs.

$$X^t \Omega X = \frac{n-1}{n-p-1} \frac{K}{K-1} \sum_{g=1}^{K} X_g^t \hat{\mathbf{e}} \hat{\mathbf{e}}^{\mathbf{t}} X_g$$

where K is the number of clusters and X_g is the subset of observations belonging to the g^{th} cluster.

Interactions

By interaction is meant an artificial factor (not measured) which reflects the interaction between at least two measured factors. For example, if we carry out treatment on a plant, and tests are carried out under two different light intensities, we will be able to include in the model an

interaction factor treatment*light which will be used to identify a possible interaction between the two factors. If there is an interaction between the two factors, we will observe a significantly larger effect on the plants when the light is strong and the treatment is of type 2 while the effect is average for weak light, treatment 2 and strong light, treatment 1 combinations.

To make a parallel with linear regression, the interactions are equivalent to the products between the continuous explanatory values although here obtaining interactions requires nothing more than simple multiplication between two variables. However, the notation used to represent the interaction between factor A and factor B is A*B.

The interactions to be used in the model can be easily defined in XLSTAT.

Nested effects

When constraints prevent us from crossing every level of one factor with every level of the other factor, nested factors can be used. We say we have a nested effect when fewer than all levels of one factor occur within each level of the other factor. An example of this might be if we want to study the effects of different machines and different operators on some output characteristic, but we can't have the operators change the machines they run. In this case, each operator is not crossed with each machine but rather only runs one machine. XLSTAT has an automatic device to find nested factors and one nested factor can be included in the model.

Balanced and unbalanced ANOVA

We talk of balanced ANOVA when for each factor (and interaction if available) the number of observations within each category is the same. When this is not true, the ANOVA is said to be unbalanced. XLSTAT can handle both cases.

Random effects

Random factors can be included in an ANOVA. When some factors are supposed to be random, XLSTAT displays the expected mean squares table.

Restricted anova

The restricted model assumptions arise when we have an interaction between a fixed factor and a random factor and assume that the sum of the random coefficients within the interaction term across each index of the fixed factor is zero. In a nutshell, the sum of the interaction effects on the levels of the fixed factor is zero.

Constraints

During the calculations, each factor is broken down into a sub-matrix containing as many columns as there are categories in the factor. Typically, this is a full disjunctive table. Nevertheless, the breakdown poses a problem: if there are g categories, the rank of this sub-matrix is not g but $g-1$. This leads to the requirement to delete one of the columns of the sub-matrix and possibly to transform the other columns. Several strategies are available depending on the interpretation we want to make afterwards:

1) **a1=0**: the parameter for the first category is null. This choice allows us force the effect of the first category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group 1.

2) **an=0**: the parameter for the last category is null. This choice allows us force the effect of the last category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group g.

3) **Sum(ai)=0**: the sum of the parameters is null. This choice forces the constant of the model to be equal to the mean of the dependent variable when the ANOVA is balanced.

4) **Sum(ni.ai)=0**: the weighted sum of the parameters is null. This choice forces the constant of the model to be equal to the mean of the dependent variable even when the ANOVA is unbalanced.

Note: even if the choice of constraint influences the values of the parameters, it has no effect on the predicted values and on the different fitting statistics, except the Type III sum of squares.

Multiple Comparisons Tests

One of the main applications of ANOVA is multiple comparisons testing whose aim is to check if the parameters for the various categories of a factor differ significantly or not. For example, in the case where four treatments are applied to plants, we want to know not only if the treatments have a significant effect, but also if the treatments have different effects.

Numerous tests have been proposed for comparing the means of categories. The majority of these tests assume that the sample is normally distributed. XLSTAT provides the main tests including:

- **Tukey's HSD test**: this test is the most used (HSD: *Honestly Significant Difference*).
- **Fisher's LSD test**: this is Student's test that tests the hypothesis that all the means for the various categories are equal (LSD: *Least Significant Difference*).
- **Bonferroni's t* test**: this test is derived from Student's test and is less reliable as it takes into account the fact that several comparisons are carried out simultaneously. Consequently, the significance level of the test is modified according to the following formula:

$$\alpha' = \frac{\alpha}{g(g-1)/2}$$

where g is the number of categories of the factor whose categories are being compared.

- **Dunn-Sidak's test**: this test is derived from Bonferroni's test. It is more reliable in some situations.

$$\alpha' = 1 - (1-\alpha)^{2/[g(g-1)]}$$

The following tests are more complex as they are based on iterative procedures where the results depend on the number of combinations remaining to be tested for each category.

- **Newman-Keuls's test (SNK)**: this test is derived from Student's test (SNK: Student Newman-Keuls), and is very often used although not very reliable.
- **Duncan's test**: this test is little used.
- **REGWQ test**: this test is among the most reliable in a majority of situations (REGW: Ryan-Einot-Gabriel-Welsch).

Another approach is possible with the **Benjamini-Hochberg** option: use this option to control the False Discovery Rate (FDR). This p-value penalization procedure is poorly conservative.

The **Games-Howell (GH)** test can be used in one-way ANOVAs when the variances lack of homogeneity. While it can be used with unequal sample sizes, it is recommended to use it when the smallest sample has 5 elements or more, otherwise it is too liberal. The **Tamhane's T2** test is more conservative, but not as powerful as the GH test.

All the above tests enable comparisons to be made between all pairs of categories and belong to the MCA test family (*Multiple Comparisons of All*, or *All-Pairwise Comparisons*).

Other tests make comparisons between all categories and a control category. These tests are called MCB tests (*Multiple Comparisons with the Best*, *Comparisons with a control*). XLSTAT offers the Dunnett test which is the most used. There are three Dunnett tests:

- **Two-tailed test**: the null hypothesis assumes equality between the category tested and the control category. The alternative hypothesis assumes the means of the two categories differ.
- **Left one-tailed test**: the null hypothesis assumes equality between the category tested and the control category. The alternative hypothesis assumes that the mean of the control category is less than the mean of the category tested.
- **Right one-tailed test**: the null hypothesis assumes equality between the category tested and the control category. The alternative hypothesis assumes that the mean of the control category is greater than the mean of the category tested.

Robust tests of mean comparison for a one-way ANOVA

In an analysis of variance, it may happen that the variances can not be assumed to be equal. In this case, the F test of the ANOVA is not robust enough to be used. XLSTAT offers two tests based on the F distribution but more robust than the classical F test.

These tests are:

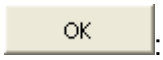
- Welch Test or Welch ANOVA (Welch, 1951). The Welch test adjusts the denominator of the F ratio so it has the same expectation as the numerator when the null hypothesis is

true, despite the heterogeneity of within-group variance. The p-value can be interpreted in the same manner as in the analysis of variance table.

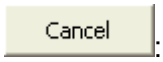
- The Brown-Forsythe test or Brown-Forsythe F-ratio (1974). This test uses a different denominator for the formula of F in the ANOVA. Instead of dividing by the mean square of the error, the mean square is adjusted using the observed variances of each group. The p-value can be interpreted in the same manner as in the analysis of variance table.

Dialog box

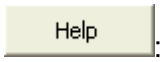
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data format :

Column: Activate this option to enter the data as columns.

Table: Activate this option to enter the data as a table (see Data selection section above for more details). A maximum of three factors is then allowed.

Data format = Column :

- **Y / Dependent variables**:
- **Quantitative**: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

- **X / Explanatory variables**:

- **Quantitative**: Activate this option to perform an ANCOVA analysis. Then select the quantitative explanatory variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

- **Qualitative**: Select the qualitative explanatory variables (the factors) in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Data format = Table :

- **Data table**: Select the data table with the response variable and the explanatory variables.
- **Number of factors**: Enter the number of factors for your analysis.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

Regression weights: Activate this option if you want to carry out a weighted least squares regression. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Groups: Activate this option if you want to group the data. Then select the data that correspond to the group to which each observation belongs.

Regression by group: enable this option if you want the calculations to be performed on each group separately. If this option is not checked, the groups will appear in different colors on the

graphs, but the ANOVA will be done on the entire data.

Options tab:

Model sub-tab:

Fixed constant: Activate this option to fix the constant of the regression model to a value you then enter (0 by default).

Tolerance: Activate this option to prevent the model from taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

Model selection: Activate this option if you want to use one of the four selection methods provided:

- **Best model**: This method lets you choose the best model from amongst all the models which can handle a number of variables varying from "Min variables" to "Max Variables". Furthermore, the user can choose several "criteria" to determine the best model.
- **Criterion**: Choose the criterion from the following list: Adjusted R^2, Mean Square of Errors (MSE), Mallows Cp, Akaike's AIC, Schwarz's SBC, Amemiya's PC.
- **Min variables**: Enter the minimum number of variables to be used in the model.
- **Max variables**: Enter the maximum number of variables to be used in the model.

Note: this method can cause long calculation times as the total number of models explored is the sum of the Cn,ks for k varying from "Min variables" to "Max variables", where Cn,k is equal to n!/[(n-k)!k !]. It is there recommended that the value of "Max variables" be increased gradually.

- **Stepwise**: The selection process starts by adding the variable with the largest contribution to the model (the criterion used is Student's t statistic). If a second variable is such that the probability associated with its t is less than the "**Probability for entry** ", it is added to the model. The same for a third variable. After the third variable is added, the impact of removing each variable present in the model after it has been added is evaluated (still using the t statistic). If the probability is greater than the "**Probability of removal** ", the variable is removed. The procedure continues until no more variables can be added or removed.

- **Forward**: The procedure is the same as for stepwise selection except that variables are only added and never removed.
- **Backward**: The procedure starts by simultaneously adding all variables. The variables are then removed from the model following the procedure used for stepwise selection.

ANOVA/ANCOVA sub-tab:

Constraints: Details on the various options are available in the description section.

- **a1 = 0**: Choose this option so that the parameter of the first category of each factor is set to 0.
- **an = 0**: Choose this option so that the parameter of the last category of each factor is set to 0.
- **Sum (ai) = 0**: for each factor, the sum of the parameters associated with the various categories is set to 0.
- **Sum (ni.ai) = 0**: for each factor, the sum of the parameters associated with the various categories weighted by their frequencies is set to 0.

Nested effects: Activate this option to include one nested effect in the model.

Random factors: Activate this option if you want to to include random factors in your model. Their impact will only be seen on the expected mean squares in the *Mixed models - Analysis Type III Sum of Squares* table.

Restricted anova: Activate this option to perform the calculation of mixed effects (fixed / random) according to the constraints imposed by the restricted anova. Their impact will only be seen on Fisher's F tests and expected mean squares from the table *Mixed models - Analysis Type III Sum of Squares*.

Covariances sub-tab:

In this tab you can choose to apply corrections for heteroscedasticity and autocorrelation. See the description section for computational details.

Validation tab:

Validation: Activate this option if you want to use a sub-sample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.

- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

Quantitative: Activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

Qualitative: Activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

Observations labels: activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

- **Check for each Y separately**: Choose this option to remove the observations with missing data in the select Y (dependent) variables, only if the Y of interest has a missing data.
- **Across all Ys**: Choose this option to remove the observations with missing data in the Y (dependent) variables, even if the Y of interest has no missing data.

Ignore missing data: Activate this option to ignore missing data. If missing data are present for the dependent variable, the corresponding observations will be predicted. If missing data are present for the explanatory variable(s) the corresponding observations are used to estimate the correlation matrix with pairwise deletion.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

General sub-tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the correlation matrix for quantitative variables (dependent or explanatory).

Effect size measures : activate this option to display the effect size measures. As part of the ANOVA, the following measures are displayed: * eta squared : $\eta^2 = \frac{SS(Factor)}{SS(Total)}$ * partial eta squared : $\eta_p^2 = \frac{SS(Factor)}{SS(Factor)+SS(Total)}$ * omega squared : $\omega^2 = \frac{SS(Factor)-df*MSE}{SS(Total)+MSE}$ * Cohen'F : $F = \sqrt{\frac{\eta_p^2}{1-\eta_p^2}}$

Multicolinearity statistics: Activate this option to display the multicolinearity statistics for all explanatory variables.

Analysis of variance: Activate this option to display the analysis of variance table.

Type I/II/III SS: Activate this option to display the Type I, Type II, and Type III sum of squares tables. Type II table is only displayed if it is different from Type III.

Press: Activate this option to calculate and display the Press (predicted residual error sum of squares) statistic.

Interpretation: Activate this option to display an automatic interpretation of the results.

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

- **Confidence intervals**: Activate this option to calculate and display the confidence intervals on the predictions.
- **Adjusted predictions**: Activate this option to calculate and display adjusted predictions in the table of predictions and residuals.
- **Influence diagnostics**: Activate this option to calculate and display the table that contains the influence statistics for each observation.

Welch and Brown-Forsythe tests: Activate this option to display the Welch and Brown-Forsythe tests (see the description section of this chapter) in the case of a one-way ANOVA.

Means sub-tab:

Means: Activate this option to compute and display the means for the categories of the main and interaction factors.

- **LS Means**: Activate this option to compute least squares means instead of observed means.
- **Standard errors**: Activate this option to display the standard errors with the means
- **Confidence intervals**: Activate this option to additionally display the confidence intervals around the means.
- **Sort**: Activate this option to display the means sorted in ascending order.
- **Multiple comparisons**: Activate this option to run multiple comparison tests (post hoc tests). Information on the multiple comparison tests is available in the description section.
- **Apply to all factors**: Activate this option to compute the selected tests for all factors.
- **Sort up**: Activate this option to sort the compared categories in increasing order, the sort criterion being their respective means. If this option is not activated, the sort is decreasing.
- **Confidence intervals**: Activate this option if you want to display the confidence intervals around the mean differences.
- **Pairwise comparisons**: Activate this option then choose the comparison methods.
- **Comparisons with a control**: Activate this option then choose the type of Dunnett test you want to carry out.
- **Choose the MSE**: Activate this option to select the mean squared error to be taken as reference for multiple comparisons. When using random factors, using the mean squared error of the model (classical case) is not appropriate. In that case, the user should choose a mean square error associated with another term in the model (usually an interaction term). If this option is enabled, a new dialog allowing you to select the variable to use will appear.

Contrasts sub-tab:

Compute contrasts: Activate this option to compute contrasts, then select the contrasts table, where there must be one column per contrast and one row for each coefficient of the model.

Test assumptions sub-tab:

These options are only available if the **Prediction and residuals** have been requested in the Outputs/General tab.

Normality test: Activate this option to run a Shapiro-Wilk test on the residuals.

Levene's test: Activate this option to run a test on the homogeneity of variances. A test is run to compare for each factor, the variance of the different categories.

Charts tab:

Regression charts: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.
- **Predictions and residuals**: Activate this option to display the following charts.

(1) Line of regression: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

(2) Explanatory variable versus standardized residuals: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

(3) Dependent variable versus standardized residuals.

(4) Predictions for the dependent variable versus the dependent variable.

(5) Bar chart of standardized residuals.

Confidence intervals: Activate this option to have confidence intervals displayed on charts (1) and (4).

Means charts: Activate this option to display the charts used to display the means of the various categories of the various factors.

- **Confidence intervals**: Activate this option to display the confidence intervals around the means on the same chart.

Summary charts: Activate this option to display summary charts that display for each factor, the comparison of the means of all categories across all dependent variables Y. **Filter Ys** allows you do decide whether the results should be displayed only for Ys for which the Fisher's F of the model is significant, or if they should be displayed for all Ys.

Results

Summary statistics: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the dependent variables (in blue) and the quantitative explanatory variables. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

Correlation matrix: This table is displayed to give you a view of the correlations between the various variables selected.

Summary of the variables selection: Where a selection method has been chosen, XLSTAT displays the selection summary. For a stepwise selection, the statistics corresponding to the different steps are displayed. Where the best model for a number of variables varying from p to q has been selected, the best model for each number or variables is displayed with the corresponding statistics and the best model for the criterion chosen is displayed in bold.

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.
- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).
- **R²**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ with } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

The R^2 is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer R^2 is to 1, the better is the model. The problem with the R^2 is that it does not take into account the number of variables used to fit the model.

- **Adjusted R²**: The adjusted determination coefficient for the model. The adjusted R^2 can be negative if the R^2 is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}$$

The adjusted R² is a correction to the R² which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W - p^*}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.
- **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows:

$$MAPE = \frac{100}{W}\sum_{i=1}^{n} w_i \left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

- **DW**: The Durbin-Watson statistic is defined by:

$$DW = \frac{\sum_{i=2}^{n}[(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2}$$

This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- **Cp**: Mallows Cp coefficient is defined by:

$$Cp = \frac{SSE}{\hat{\sigma}} + 2p^* - W$$

where SSE is the sum of the squares of the errors for the model with p explanatory variables and $\hat{\sigma}$ is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to p*, the less the model is biased.

- **AIC**: Akaike's Information Criterion is defined by:

$$AIC = W \ln(\frac{SSE}{W}) + 2p^*$$

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz's Bayesian Criterion is defined by:

$$SBC = W \ln(\frac{SSE}{W}) + \ln(W)p^*$$

This criterion, proposed by Schwarz (1978) is similar to the AIC, and the aim is to minimize it.

- **PC**: Amemiya's Prediction Criterion is defined by:

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}$$

This criterion, proposed by Amemiya (1980) is used, like the adjusted R² to take account of the parsimony of the model.

- **Press**: The Press (predicted residual error sum of squares) statistic is only displayed if the corresponding option has been activated in the dialog box. It is defined by:

$$Press = \sum_{i=1}^{n} w_i(y_i - \hat{y}_{i(-i)})^2$$

where $\hat{y}_{i(-i)}$is the prediction for observation i when the latter is not used for estimating parameters. We then get:

$$\text{Press } RMSE = \sqrt{\frac{Press}{W - p^*}}$$

The Press RMSE can then be compared to the RMSE. A large difference between the two shows that the model is sensitive to the presence or absence of certain observations in the model.

- **Q²**: This statistic, also known as the cross-validated R², is only displayed if the Press option has been activated in the dialog box. It is defined by:

$$Q^2 = 1 - \frac{\text{Press}}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

This gives the proportion of the total variance that is explained by the explanatory variables when the predictions are computed when the corresponding observation is not in the model. A large difference between the Q² and the R² shows that the model is sensitive to the presence or absence of certain observations in the model.

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

If the Type I/II/III SS (SS: Sum of Squares) is activated, the corresponding tables are displayed.

The table of **Type I SS** values is used to visualize the influence that progressively adding explanatory variables has on the fitting of the model, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. The sums of squares in the Type I table always add up to the model SS. Note: the order in which the variables are selected in the model influences the values obtained.

The table of **Type II SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained.

The table of **Type III SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, expect those were the effect is present (interactions), as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the

probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. Type II and Type III are identical if there are no interactions or if the design is balanced.

The **parameters of the model** table displays the estimate of the parameters, the corresponding standard error, the Student's t, the corresponding probability, as well as the confidence interval

The **equation of the model** is then displayed to make it easier to read or re-use the model.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the residuals, the confidence intervals together with the adjusted prediction if the corresponding options have been activated in the dialog box. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the uncertainty being larger. If validation data have been selected, they are displayed at the end of the table.

The **charts** which follow show the results mentioned above. A chart shows the standardized residuals as a function of the explanatory variable. In principle, the residuals should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a problem with the model.

The **three charts** displayed next show respectively the evolution of the standardized residuals as a function of the dependent variable, the distance between the predictions and the observations (for an ideal model, the points would all be on the bisector), and the standardized residuals on a bar chart. The last chart quickly shows if an abnormal number of values are outside the interval]-2, 2[given that the latter, assuming that the sample is normally distributed, should contain about 95% of the data.

The table of **influence diagnostics** displays for each observation, its weight, the corresponding residual, the standardized residual (divided by the $RMSE$), the studentized residual, the deleted residual, the studentized deleted residual, the centered leverage, the Mahalanobis distance, the Cook's D, the CovRatio, the DFFit, the standardized DFFit, the DFBetas (one per model coefficient) and the standardized DFBetas.

Four charts are then displayed to make possible an easy identification of the observations which influence on the predictions or on the coefficients make necessary a special investigation.

If you have selected the data to be used for calculating **predictions on new observations**, the corresponding table is displayed next.

If the **tests on assumptions** have been requested, the corresponding results are then displayed.

If **multiple comparison tests** have been requested, the corresponding results are then displayed.

When a one-way ANOVA is applied and the corresponding option is enabled, the results of the Welch and Brown-Forsythe tests are displayed. The associated statistics, the degrees of freedom and the p-values are displayed.

If several dependent variables have been selected and if the multiple comparisons option has been activated, a table showing the means for each category of each factor and across all Ys is displayed. The cells of the table are colored using a spectrum scale from blue to red. If there are more than 10 categories only the 5 lowest and 5 highest means are colored. A chart allows visualizing the same results.

A second chart allows to visualize estimated means with multiple comparisons grouping letters. This second chart only displays significant dependent variables on the factor in question according to the Type III SS ANOVA table.

Example

A tutorial on one-way ANOVA and multiple comparisons tests is available on the XLSTAT Help Center:

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

A tutorial on two-way ANOVA is available on the XLSTAT Help Center:

http://www.xlstat.com/demo-ano2.htm

References

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. & Robotten M. (1997). Analysis of Variance for Sensory Data, John Wiley & 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.

ANCOVA

Use this module to model a quantitative dependent variable by using quantitative and qualitative dependent variables as part of a linear model.

In this section:

Description

ANCOVA (ANalysis of COVAriance) can be seen as a mix of ANOVA and linear regression as the dependent variable is of the same type, the model is linear and the hypotheses are identical. In reality it is more correct to consider ANOVA and linear regression as special cases of ANCOVA.

If p is the number of quantitative variables, and q the number of factors (the qualitative variables including the interactions between qualitative variables), the ANCOVA model is written as follows:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{j=1}^{q} \beta_{k(i,j),j} + \epsilon_i \quad (1)$$

where y_i is the value observed for the dependent variable for observation i, x_{ij} is the value taken by quantitative variable j for observation i, $k(i,j)$ is the index of the category of factor j for observation i and ϵ_i is the error of the model.

The estimator of the β coefficients and of their covariance matrix are given by:

$$\hat{\beta} = (X^t X)^{-1} X^t Y \quad (2)$$

and

$$Var(\hat{\beta}) = \hat{\sigma}^2 (X^t X)^{-1} \quad (3)$$

To use the various tests proposed in the results ANCOVA, it is recommended to check retrospectively that the underlying hypotheses have been correctly verified. The normality of the residuals can be checked by analyzing certain charts, by using a normality test, or by running a Levene's test to test if the variances for the different groups are homogeneous. The

independence of the residuals can be checked by analyzing certain charts or by using the Durbin Watson test.

Interactions between quantitative variables and factors

One of the features of ANCOVA is to enable interactions between quantitative variables and factors to be taken into account. The main application is to test if the level of a factor (a qualitative variable) has an influence on the coefficient (often called slope in this context) of a quantitative variable. Comparison tests are used to test if the slopes corresponding to the various levels of a factor differ significantly or not. A model with one quantitative variable and a factor with interaction is written:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_{k(i,1),1} + \beta_{k(i,1),2} x_{i1} + \epsilon_i. \quad (4)$$

This can be simplified by setting

$$\gamma_{k(i,1),1} = \beta_1 + \beta_{k(i,1),2} \quad (5)$$

hence we get

$$y_i = \beta_0 + \beta_{k(i,1),1} + \gamma_{k(i,1),1} x_{i1} + \epsilon_i. \quad (6)$$

The comparison of the γ parameters are used to test if the factor has an affect on the slope.

Correcting for Heteroscedasticity and Autocorrelation

Homoscedasticity and independence of the error terms are key hypotheses in linear regression and ANOVA where it is assumed that the variances of the error terms are independent and identically distributed and normally distributed. When these assumptions are not possible to keep (a Durbin Watson or White test available in the Time series menu allow to challenge these hypotheses), a consequence is that the covariance matrix cannot be estimated using the classical formula, and the variance of the parameters corresponding to the coefficients of the linear model can be wrong and their confidence intervals as well. A predictor could be said to be significant (or respectively not) while being the opposite. XLSTAT allows to correct for heteroscedasticity and autocorrelation that can arise, especially in time series.

For what concerns heteroscedasticity, White (1980) followed by several authors has explored ways to correct the classical estimate of the covariances using residuals and centered leverages obtained from the linear regression computations (see MacKinnon (1985) and Zeileis (2006) for a review). When the assumptions of classical linear regression do not hold, while the estimators of the coefficients are unchanged, the simplified writing of the covariance matrix (see equation 3) of the β parameters is not possible, and we must revert to the general expression:

$$Var(\beta) = (X^t X)^{-1} (X^t \Omega X) (X^t X)^{-1}. \quad (7)$$

Equation (7) is equivalent to equation (3) when

$$\Omega = \hat{\sigma}^2 I. \quad (8)$$

Let ω_i be the diagonal elements of W. The different heteroscedasticity coefficients (HC) estimators for the ω_i are given by:

$$HC0: \quad \omega_i = \hat{e_i}^2$$
$$HC1: \quad \omega_i = \hat{e_i}^2 \frac{n}{(n-p-1)}$$
$$HC2: \quad \omega_i = \frac{\hat{e_i}^2}{(1-h_i)}$$
$$HC3: \quad \omega_i = \frac{\hat{e_i}^2}{(1-h_i)^2}$$
$$HC4: \quad \omega_i = \frac{\hat{e_i}^2}{(1-h_i)^{\delta_i}} \text{ with } \delta_i = min(4, h_i/\bar{h})$$

where $\hat{e_i}$ is the residual, and h_i the centered leverage for the ith observation, and p is the number of predictors.

Newey and West (1987) suggested an estimator that corrects for both autocorrelation and heteroscedasticity, but the lag must be known from the user (the descriptive analysis of time series or ARIMA functions of XLSTAT can be used for that). For lag=0 (no autocorrelation) we have:

$$X^t \Omega X = X^t \Omega_0 X = \frac{n}{n-p-1} \sum_{i=1}^{n} \hat{e}_i^2 x_i^t x_i$$

where x_i is the vector of the predictors (including a 1 for the intercept of the model) for the ith observation. For lag m (m>0), we have:

$$X^t \Omega X = X^t \Omega_0 X + \frac{n}{n-p-1} \sum_{l=1}^{m} \sum_{t=l+1}^{n} \hat{e}_t^2 \hat{e}_{t-l}^2 (x_t^t x_{t-l} - x_{t-l}^t x_t).$$

The unadjusted version of the Newey West estimator corresponds to the same without the $n/(n-p-1)$ adjustment factor.

The **Clusters** option makes it possible to correct the problem of heteroscedasticity in the case where the variances are considered to be equal only within given specific clusters. When this option is selected, you must then select the data indicating to which cluster each observation belongs.

$$X^t \Omega X = \frac{n-1}{n-p-1} \frac{K}{K-1} \sum_{g=1}^{K} X_g^t \hat{\mathbf{e}} \hat{\mathbf{e}}^{\mathbf{t}} X_g$$

where K is the number of clusters and X_g is the subset of observations belonging to the g^{th} cluster.

Random effects

Random factors can be included in an ANOVA. When some factors are supposed to be random, XLSTAT displays the expected mean squares table.

Constraints

During the calculations, each factor is broken down into a sub-matrix containing as many columns as there are categories in the factor. Typically, this is a full disjunctive table. Nevertheless, the breakdown poses a problem: if there are g categories, the rank of this sub-matrix is not g but $g-1$. This leads to the requirement to delete one of the columns of the sub-matrix and possibly to transform the other columns. Several strategies are available depending on the interpretation we want to make afterwards:

1) **a1=0**: the parameter for the first category is null. This choice allows us force the effect of the first category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group 1.

2) **an=0**: the parameter for the last category is null. This choice allows us force the effect of the last category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group g.

3) **Sum(ai)=0**: the sum of the parameters is null. This choice forces the constant of the model to be equal to the mean of the dependent variable when the ANOVA is balanced.

4) **Sum(ni.ai)=0**: the weighted sum of the parameters is null. This choice forces the constant of the model to be equal to the mean of the dependent variable even when the ANOVA is unbalanced.

Note: even if the choice of constraint influences the values of the parameters, it has no effect on the predicted values and on the different fitting statistics, except the Type III sum of squares.

Multiple Comparisons Tests

One of the main applications of ANOVA is multiple comparisons testing whose aim is to check if the parameters for the various categories of a factor differ significantly or not. For example, in the case where four treatments are applied to plants, we want to know not only if the treatments have a significant effect, but also if the treatments have different effects.

Numerous tests have been proposed for comparing the means of categories. The majority of these tests assume that the sample is normally distributed. XLSTAT provides the main tests including:

- **Tukey's HSD test**: this test is the most used (HSD: *Honestly Significant Difference*).
- **Fisher's LSD test**: this is Student's test that tests the hypothesis that all the means for the various categories are equal (LSD: *Least Significant Difference*).
- **Bonferroni's t* test**: this test is derived from Student's test and is less reliable as it takes into account the fact that several comparisons are carried out simultaneously. Consequently, the significance level of the test is modified according to the following formula:

$$\alpha' = \frac{\alpha}{g(g-1)/2},$$

where g is the number of categories of the factor whose categories are being compared.

- **Dunn-Sidak's test**: this test is derived from Bonferroni's test. It is more reliable in some situations.

$$\alpha' = 1 - (1 - \alpha)^{2/[g(g-1)]}.$$

The following tests are more complex as they are based on iterative procedures where the results depend on the number of combinations remaining to be tested for each category.

- **Newman-Keuls's test (SNK)**: this test is derived from Student's test (SNK: Student Newman-Keuls), and is very often used although not very reliable.
- **Duncan's test**: this test is little used.
- **REGWQ test**: this test is among the most reliable in a majority of situations (REGW: Ryan-Einot-Gabriel-Welsch).

Another approach is possible with the **Benjamini-Hochberg** option: use this option to control the False Discovery Rate (FDR). This p-value penalization procedure is poorly conservative.

The **Games-Howell (GH)** test can be used in one-way ANOVAs when the variances lack of homogeneity. While it can be used with unequal sample sizes, it is recommended to use it when the smallest sample has 5 elements or more, otherwise it is too liberal. The **Tamhane's T2** test is more conservative, but not as powerful as the GH test.

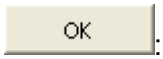
All the above tests enable comparisons to be made between all pairs of categories and belong to the MCA test family (*Multiple Comparisons of All*, or *All-Pairwise Comparisons*).

Other tests make comparisons between all categories and a control category. These tests are called MCB tests (*Multiple Comparisons with the Best*, *Comparisons with a control*). XLSTAT offers the Dunnett test which is the most used. There are three Dunnett tests:

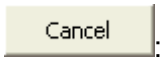
- **Two-tailed test**: the null hypothesis assumes equality between the category tested and the control category. The alternative hypothesis assumes the means of the two categories differ.
- **Left one-tailed test**: the null hypothesis assumes equality between the category tested and the control category. The alternative hypothesis assumes that the mean of the control category is less than the mean of the category tested.
- **Right one-tailed test**: the null hypothesis assumes equality between the category tested and the control category. The alternative hypothesis assumes that the mean of the control category is greater than the mean of the category tested.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Dependent variables:

Quantitative: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

X / Explanatory variables:

Quantitative: Select the quantitative explanatory variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Select the qualitative explanatory variables (the factors) in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

Regression weights: Activate this option if you want to carry out a weighted least squares regression. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Groups: Activate this option if you want to group the data. Then select the data that correspond to the group to which each observation belongs.

Options tab:

Model sub-tab:

Fixed constant: Activate this option to fix the constant of the regression model to a value you then enter (0 by default).

Tolerance: Activate this option to prevent the model from taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

Model selection: Activate this option if you want to use one of the four selection methods provided:

- **Best model**: This method lets you choose the best model from amongst all the models which can handle a number of variables varying from "Min variables" to "Max Variables".

Furthermore, the user can choose several "criteria" to determine the best model.

- **Criterion**: Choose the criterion from the following list: Adjusted R², Mean Square of Errors (MSE), Mallows Cp, Akaike's AIC, Schwarz's SBC, Amemiya's PC.
- **Min variables**: Enter the minimum number of variables to be used in the model.
- **Max variables**: Enter the maximum number of variables to be used in the model.

Note: this method can cause long calculation times as the total number of models explored is the sum of the $C_{n,k}$s for k varying from "Min variables" to "Max variables", where $C_{n,k}$ is equal to $n!/[(n-k)!k!]$. It is there recommended that the value of "Max variables" be increased gradually.

- **Stepwise**: The selection process starts by adding the variable with the largest contribution to the model (the criterion used is Student's t statistic). If a second variable is such that the probability associated with its t is less than the "**Probability for entry** ", it is added to the model. The same for a third variable. After the third variable is added, the impact of removing each variable present in the model after it has been added is evaluated (still using the t statistic). If the probability is greater than the "**Probability of removal** ", the variable is removed. The procedure continues until no more variables can be added or removed.
- **Forward**: The procedure is the same as for stepwise selection except that variables are only added and never removed.
- **Backward**: The procedure starts by simultaneously adding all variables. The variables are then removed from the model following the procedure used for stepwise selection.

ANOVA/ANCOVA sub-tab:

Constraints: Details on the various options are available in the description section of the ANOVA.

- **a1 = 0**: Choose this option so that the parameter of the first category of each factor is set to 0.
- **an = 0**: Choose this option so that the parameter of the last category of each factor is set to 0.
- **Sum (ai) = 0**: for each factor, the sum of the parameters associated with the various categories is set to 0.
- **Sum (ni.ai) = 0**: for each factor, the sum of the parameters associated with the various categories weighted by their frequencies is set to 0.

Nested effects: Activate this option to include one nested effect in the model.

Covariances sub-tab:

In this tab you can choose to apply corrections for heteroscedasticity and autocorrelation. See the *description* section for computational details.

Validation tab:

Validation: Activate this option if you want to use a sub-sample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

Quantitative: Activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

Qualitative: Activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

Observations labels: activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

- **Check for each Y separately**: Choose this option to remove the observations with missing data in the select Y (dependent) variables, only if the Y of interest has a missing data.
- **Across all Ys**: Choose this option to remove the observations with missing data in the Y (dependent) variables, even if the Y of interest has no missing data.

Ignore missing data: Activate this option to ignore missing data. If missing data are present for the dependent variable, the corresponding observations will be predicted. If missing data are present for the explanatory variable(s) the corresponding observations are used to estimate the correlation matrix with pairwise deletion.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Ouputs tab:

General sub-tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the correlation matrix for quantitative variables (dependent or explanatory).

Multicolinearity statistics: Activate this option to display the multicolinearity statistics for all explanatory variables.

Analysis of variance: Activate this option to display the analysis of variance table.

Type I/II/III SS: Activate this option to display the Type I, Type II, and Type III sum of squares tables. Type II table is only displayed if it is different from Type III.

Press: Activate this option to calculate and display the Press (predicted residual error sum of squares) statistic.

Interpretation: Activate this option to display an automatic interpretation of the results.

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

- **X**: Activate this option to display the explanatory variables in the predictions and residuals table.

- **Confidence intervals**: Activate this option to calculate and display the confidence intervals on the predictions.
- **Adjusted predictions**: Activate this option to calculate and display adjusted predictions in the table of predictions and residuals.
- **Influence diagnostics**: Activate this option to calculate and display the table that contains the influence statistics for each observation.

Welch and Brown-Forsythe tests: Activate this option to display the Welch and Brown-Forsythe tests (see the description section of this chapter) in the case of a one-way ANOVA.

Means sub-tab:

Means: Activate this option to compute and display the means for the categories of the main and interaction factors.

- **LS Means**: Activate this option to compute least squares means instead of observed means.
- **Standard errors**: Activate this option to display the standard errors with the means
- **Confidence intervals**: Activate this option to additionally display the confidence intervals around the means.
- **Sort**: Activate this option to display the means sorted in ascending order.
- **Multiple comparisons**: Activate this option to run multiple comparison tests (post hoc tests). Information on the multiple comparison tests is available in the description section.
- **Apply to all factors**: Activate this option to compute the selected tests for all factors.
- **Sort up**: Activate this option to sort the compared categories in increasing order, the sort criterion being their respective means. If this option is not activated, the sort is decreasing.
- **Confidence intervals**: Activate this option if you want to display the confidence intervals around the mean differences.
- **Pairwise comparisons**: Activate this option then choose the comparison methods.
- **Comparisons with a control**: Activate this option then choose the type of Dunnett test you want to carry out.
- **Choose the MSE**: Activate this option to select the mean squared error to be taken as reference for multiple comparisons. When using random factors, using the mean squared error of the model (classical case) is not appropriate. In that case, the user should choose a mean square error associated with another term in the model (usually an interaction term). If this option is enabled, a new dialog allowing you to select the variable to use will appear.
- **Comparison of slopes**: Activate this option to compare the interaction slopes between the quantitative and qualitative variables (see the description section on this subject).

Contrasts sub-tab:

Compute contrasts: Activate this option to compute contrasts, then select the contrasts table, where there must be one column per contrast and one row for each coefficient of the model.

Test assumptions sub-tab:

These options are only available if the **Prediction and residuals** have been requested in the Outputs/General tab.

Normality test: Activate this option to run a Shapiro-Wilk test on the residuals.

Levene's test: Activate this option to run a test on the homogeneity of variances. A test is run to compare for each factor, the variance of the different categories.

Charts tab:

Regression charts: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.
- **Predictions and residuals**: Activate this option to display the following charts.

(1) Line of regression: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

(2) Explanatory variable versus standardized residuals: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

(3) Dependent variable versus standardized residuals.

(4) Predictions for the dependent variable versus the dependent variable.

(5) Bar chart of standardized residuals.

Confidence intervals: Activate this option to have confidence intervals displayed on charts (1) and (4).

Means charts: Activate this option to display the charts used to display the means of the various categories of the various factors.

- **Confidence intervals**: Activate this option to display the confidence intervals around the means on the same chart.

Summary charts: Activate this option to display summary charts that display for each factor, the comparison of the means of all categories across all dependent variables Y. **Filter Ys** allows you do decide whether the results should be displayed only for Ys for which the Fisher's F of the model is significant, or if they should be displayed for all Ys.

Results

Summary statistics: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the dependent variables (in blue) and the quantitative explanatory variables. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

Correlation matrix: This table is displayed to give you a view of the correlations between the various variables selected.

Summary of the variables selection: Where a selection method has been chosen, XLSTAT displays the selection summary. For a stepwise selection, the statistics corresponding to the different steps are displayed. Where the best model for a number of variables varying from p to q has been selected, the best model for each number or variables is displayed with the corresponding statistics and the best model for the criterion chosen is displayed in bold.

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.
- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).
- **R²**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ with } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

The R² is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer R² is to 1, the better is the model. The problem with the R² is that it does not take into account the number of variables used to fit the model.

- **Adjusted R²**: The adjusted determination coefficient for the model. The adjusted R² can be negative if the R² is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}.$$

The adjusted R² is a correction to the R² which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W - p^*} \sum_{i=1}^{x} w_i (y_i - \hat{y}_i)^2.$$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.
- **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows:

$$MAPE = \frac{100}{W} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|.$$

- **DW**: The Durbin-Watson statistic is defined by:

$$DW = \frac{\sum_{i=2}^{x} [(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{x} w_i (y_i - \hat{y}_i)^2}.$$

This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- **Cp**: Mallows Cp coefficient is defined by:

$$Cp = \frac{SSE}{\hat{\sigma}} + 2p^* - W,$$

where SSE is the sum of the squares of the errors for the model with p explanatory variables and $\hat{\sigma}$ is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to p*, the less the model is biased.

- **AIC**: Akaike's Information Criterion is defined by:

$$AIC = W \ln\left(\frac{SSE}{W}\right) + 2p^*.$$

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz's Bayesian Criterion is defined by:

$$SBC = W \ln\left(\frac{SSE}{W}\right) + \ln(W) p^*.$$

This criterion, proposed by Schwarz (1978) is similar to the AIC, and the aim is to minimize it.

- **PC**: Amemiya's Prediction Criterion is defined by:

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}.$$

This criterion, proposed by Amemiya (1980) is used, like the adjusted R² to take account of the parsimony of the model.

- **Press**: The Press (predicted residual error sum of squares) statistic is only displayed if the corresponding option has been activated in the dialog box. It is defined by:

$$Press = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i(-i)})^2,$$

where $\hat{y}_{i(-i)}$ is the prediction for observation i when the latter is not used for estimating parameters. We then get:

$$\text{Press } RMSE = \sqrt{\frac{Press}{W - p^*}}.$$

The Press RMSE can then be compared to the RMSE. A large difference between the two shows that the model is sensitive to the presence or absence of certain observations in the model.

- **Q²**: This statistic, also known as the cross-validated R², is only displayed if the Press option has been activated in the dialog box. It is defined by:

$$Q^2 = 1 - \frac{\text{Press}}{\sum_{i=1}^{x} (y_i - \overline{y})^2}.$$

This gives the proportion of the total variance that is explained by the explanatory variables when the predictions are computed when the corresponding observation is not in the model. A large difference between the Q² and the R² shows that the model is sensitive to the presence or absence of certain observations in the model.

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

If the Type I/II/III SS (SS: Sum of Squares) is activated, the corresponding tables are displayed.

The table of **Type I SS** values is used to visualize the influence that progressively adding explanatory variables has on the fitting of the model, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. The sums of squares in the Type I table always add up to the model SS. Note: the order in which the variables are selected in the model influences the values obtained.

The table of **Type II SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, as regards the sum of

the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained.

The table of **Type III SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, expect those were the effect is present (interactions), as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. Type II and Type III are identical if there are no interactions or if the design is balanced.

The **parameters of the model** table displays the estimate of the parameters, the corresponding standard error, the Student’s t, the corresponding probability, as well as the confidence interval

The **equation of the model** is then displayed to make it easier to read or re-use the model.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the residuals, the confidence intervals together with the adjusted prediction if the corresponding options have been activated in the dialog box. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the uncertainty being larger. If validation data have been selected, they are displayed at the end of the table.

The **charts** which follow show the results mentioned above. If there is only one explanatory variable and one factor in the model, the first chart displayed shows the observed values and the regression line for each category of the factor. The second chart shows the standardized residuals as a function of the explanatory variable. In principle, the residuals should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a problem with the model.

The **three charts** displayed next show respectively the evolution of the standardized residuals as a function of the dependent variable, the distance between the predictions and the observations (for an ideal model, the points would all be on the bisector), and the standardized residuals on a bar chart. The last chart quickly shows if an abnormal number of values are outside the interval]-2, 2[given that the latter, assuming that the sample is normally distributed, should contain about 95% of the data.

The table of **influence diagnostics** displays for each observation, its weight, the corresponding residual, the standardized residual (divided by the $RMSE$), the studentized residual, the deleted residual, the studentized deleted residual, the centered leverage, the Mahalanobis distance, the Cook's D, the CovRatio, the DFFit, the standardized DFFit, the DFBetas (one per model coefficient) and the standardized DFBetas.

Four charts are then displayed to make possible an easy identification of the observations which influence on the predictions or on the coefficients make necessary a special investigation.

If you have selected the data to be used for calculating **predictions on new observations**, the corresponding table is displayed next.

If the **tests on assumptions** have been requested, the corresponding results are then displayed.

If **multiple comparison tests** have been requested, the corresponding results are then displayed.

If several dependent variables have been selected and if the multiple comparisons option has been activated, a table showing the means for each category of each factor and across all Ys is displayed. The cells of the table are colored using a spectrum scale from blue to red. If there are more than 10 categories only the 5 lowest and 5 highest means are colored. A chart allows visualizing the same results.

A second chart allows to visualize estimated means with multiple comparisons grouping letters. This second chart only displays significant dependent variables on the factor in question according to the Type III SS ANOVA table.

Example

A tutorial on ANCOVA is available on the XLSTAT Help Center:

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

References

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 & 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.

Repeated Measures ANOVA

Use this tool to carry out Repeated Measures ANOVA (ANalysis Of VAriance). The advanced options enable you to choose the constraints on the model and to take account of interactions between the factors. Multiple comparison tests can be calculated.

XLSTAT proposes two ways for handling repeated measures ANOVA. The classical way uses least squares estimation (LS) that is based on the same model as the classical ANOVA and the alternative way that is based on the maximum likelihood estimation (REML and ML).

This chapter is devoted to the first method. For details on the second method, please read the chapter on mixed models.

In this section:

Description

Repeated measures Analysis of Variance (ANOVA) uses the same conceptual framework as classical ANOVA. The main difference comes from the nature of the explanatory variables. The exploratory variable is measured at different time or repetition. In ANOVA, explanatory variables are often called factors.

If p is the number of factors, the ANOVA model is written as follows:

$$y_i^t = \beta_0 + \sum_{j=1}^{p} \beta_{k(i,j),j} + \epsilon_i \quad (1)$$

where y_i^t is the value observed for the dependent variable for observation i for measure t, $k(i,j)$ is the index of the category of factor j for observation i, and ϵ_i is the error of the model.

The hypotheses used in ANOVA are identical to those used in linear regression: the errors ϵ_i follow the same normal distribution $N(0, s)$ and are independent.

However, other hypotheses are necessary in the case of repeated measures ANOVA. As measures are taken from the same subjects at different times, the repetitions are correlated. In repeated measures ANOVA we assume that the covariance matrix between the ys is spherical (for example, compound symmetry is a spherical shape). We can drop this hypothesis when using the maximum likelihood based approach.

The principle of repeated measures ANOVA is simple. For each measure, a classical ANOVA model is estimated, then the sphericity of the covariance matrix between measures is tested using Mauchly's test, Greenhouse-Geisser epsilon or Huynt-Feldt epsilon. If the sphericity hypothesis is not rejected, between- and within-subject effects can be tested.

Interactions

By interaction is meant an artificial factor (not measured) which reflects the interaction between at least two measured factors. For example, if we carry out treatment on a plant, and tests are carried out under two different light intensities, we will be able to include in the model an interaction factor treatment*light which will be used to identify a possible interaction between the two factors. If there is an interaction between the two factors, we will observe a significantly larger effect on the plants when the light is strong and the treatment is of type 2 while the effect is average for weak light, treatment 2 and strong light, treatment 1 combinations.

To make a parallel with linear regression, the interactions are equivalent to the products between the continuous explanatory values although here obtaining interactions requires nothing more than simple multiplication between two variables. However, the notation used to represent the interaction between factor A and factor B is A*B.

The interactions to be used in the model can be easily defined in XLSTAT.

Nested effects

When constraints prevent us from crossing every level of one factor with every level of the other factor, nested factors can be used. We say we have a nested effect when fewer than all levels of one factor occur within each level of the other factor. An example of this might be if we want to study the effects of different machines and different operators on some output characteristic, but we can't have the operators change the machines they run. In this case, each operator is not crossed with each machine but rather only runs one machine.

XLSTAT has an automatic device to find nested factors and one nested factor can be included in the model.

Constraints

During the calculations, each factor is broken down into a sub-matrix containing as many columns as there are categories in the factor. Typically, this is a full disjunctive table. Nevertheless, the breakdown poses a problem: if there are g categories, the rank of this sub-matrix is not g but $g-1$. This leads to the requirement to delete one of the columns of the sub-matrix and possibly to transform the other columns. Several strategies are available depending on the interpretation we want to make afterwards:

1) **a1=0**: The parameter for the first category is null. This choice allows us to force the effect of the first category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group 1.

2) **an=0**: The parameter for the last category is null. This choice allows us to force the effect of the last category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group g.

Note: Even if the choice of constraint influences the values of the parameters, it has no effect on the predicted values and on the different fitting statistics.

Multiple Comparisons Tests

One of the main applications of ANOVA is multiple comparisons testing whose aim is to check if the parameters for the various categories of a factor differ significantly or not. For example, in the case where four treatments are applied to plants, we want to know not only if the treatments have a significant effect, but also if the treatments have different effects.

Numerous tests have been proposed for comparing the means of categories. Most of these tests assume that the sample is normally distributed. XLSTAT provides the main tests including:

Tukey's HSD test: This test is the most used (HSD: *Honestly Significant Difference*).

Fisher's LSD test: This is Student's test that tests the hypothesis that all the means for the various categories are equal (LSD: *Least Significant Difference*).

Bonferroni's t* test: This test is derived from Student's test and is less reliable as it considers the fact that several comparisons are carried out simultaneously. Consequently, the significance level of the test is modified according to the following formula:

$$\alpha' = \frac{\alpha}{g(g-1)/2},$$

where g is the number of categories of the factor whose categories are being compared.

Dunn-Sidak's test: this test is derived from Bonferroni's test. It is more reliable in some situations.

$$\alpha' = 1 - (1-\alpha)^{2/[g(g-1)]}.$$

The following tests are more complex as they are based on iterative procedures where the results depend on the number of combinations remaining to be tested for each category.

Newman-Keuls's test (SNK): This test is derived from Student's test (SNK: Student Newman-Keuls), and is very often used although not very reliable.

Duncan's test: This test is not widely used.

REGWQ test: This test is among the most reliable in most situations (REGW: Ryan-Einot-Gabriel-Welsch).

Another approach is possible with the **Benjamini-Hochberg** option: use this option to control the False Discovery Rate (FDR). This p-value penalization procedure is poorly conservative.

The **Games-Howell (GH)** test can be used in one-way ANOVAs when the variances lack of homogeneity. While it can be used with unequal sample sizes, it is recommended to use it when the smallest sample has 5 elements or more, otherwise it is too liberal. The **Tamhane's T2** test is more conservative, but not as powerful as the GH test.

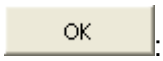
All the above tests enable comparisons to be made between all pairs of categories and belong to the MCA test family (*Multiple Comparisons of All*, or *All-Pairwise Comparisons*).

Other tests make comparisons between all categories and a control category. These tests are called MCB tests (*Multiple Comparisons with the Best*, *Comparisons with a control*). XLSTAT offers the Dunnett test, which is the most used. There are three Dunnett tests:

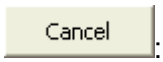
- **Two-tailed test**: The null hypothesis assumes equality between the category tested and the control category. The alternative hypothesis assumes the means of the two categories differ.
- **Left one-tailed test**: The null hypothesis assumes equality between the category tested and the control category. The alternative hypothesis assumes that the mean of the control category is greater than the mean of the category tested.
- **Right one-tailed test**: The null hypothesis assumes equality between the category tested and the control category. The alternative hypothesis assumes that the mean of the control category is less than the mean of the category tested.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Dependent variables:

Quantitative: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

One column for all repetitions: Activate this option if your dependent variable is organized in only one column. In that case, you must select as explanatory variables, one variable with the

name of the repetition and another variable with the name of the subject. For more details on that format, please see the chapter on mixed models.

One column per repetition: Activate this option if your dependent variable has T columns for T repetitions. In that case, when you select the factors, a factor called repetition and a factor called subject will appear.

X / Explanatory variables:

Qualitative: Select the qualitative explanatory variables (the factors) in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Quantitative: Select one or more quantitative variables on the Excel sheet. If the variable labels have been selected, please check the option "Variable labels" is activated. When no qualitative variables are selected, then it is a repeated measure of linear regression. If qualitative and quantitative variables are selected then it is a repeated measures ANCOVA. If no explanatory variables are selected, it is a one-way repeated measures ANOVA.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

Regression weights: Activate this option if you want to carry out a weighted least squares regression. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Fixed constant: Activate this option to fix the constant of the regression model to a value you then enter (0 by default).

Tolerance: Activate this option to prevent the OLS regression calculation algorithm considering variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Nested effects: Activate this option to include one nested effect in the model.

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

Estimation method: Three methods are available. The first one is the classical method based on least squares and noted LS. The two other two methods are based on maximum likelihood and are developed in the developed in the context of the help on mixed models (the outputs are then completely different).

Constraints: Details on the various options are available in the description section.

- **a1 = 0**: Choose this option so that the parameter of the first category of each factor is set to 0.
- **an = 0**: Choose this option so that the parameter of the last category of each factor is set to 0.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Analysis of variance: Activate this option to display the analysis of variance table for each repetition t.

Type I/III SS: Activate this option to display the Type I and Type III sum of squares tables for each ANOVA associated to repetition t.

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Test for Within subject effects: Activate this option to display tests for within-subjects tests for within-subject effects.

Test for between subject effects: Activate this option to display tests for between-subjects effects.

Mauchly's Sphericity Test: Activate this option to display Mauchly's sphericity test.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

Multiple comparisons:

Information on the multiple comparison tests is available in the description section.

Apply to all factors: Activate this option to compute the selected tests for all factors.

Use least squares means: Activate this option to compare the means using their least squares estimators (obtained from the parameters of the model). If this option is not activated, the means are computed using their estimation based on the data.

Sort up: Activate this option to sort the compared categories in increasing order, the sort criterion being their respective means. If this option is not activated, the sort decreases.

Pairwise comparisons: Activate this option then choose the comparison methods.

Comparisons with a control: Activate this option then choose the type of Dunnett test you want to carry out.

Charts tab:

Regression charts: Activate this option to display regression chart:

- **Predictions and residuals**: Activate this option to display the following charts.

  (1) Line of regression: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

  (2) Explanatory variable versus standardized residuals: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

  (3) Dependent variable versus standardized residuals.

  (4) Predictions for the dependent variable versus the dependent variable.

  (5) Bar chart of standardized residuals.

Means charts: Activate this option to display the charts used to display the means of the various categories of the various factors.

Factors and interactions dialog box

Once the first dialog box disappears, a second one appears, to allow you to clarify the belonging of each factor. It is necessary to select the fixed factors (fixed effects), a repeated factor and a subject factor. If you selected the *one column per repetition* layout, then a factor called repetition and a factor called subject are displayed and must respectively be selected as repeated and subject factors.

Results

Summary statistics: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the dependent variables (in blue) and the quantitative explanatory variables. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

Then for each repetition, we have:

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.
- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).
- **R²**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by: $$ R^ { 2} = \frac { \sum _ { i = 1} ^ { n } w _ { i } (\hat {y} _ { i } - \bar { y } _ { i }) ^ { 2} } { \sum _ { i = 1} ^ { n } w _ { i } (y _ { i } - \bar { y } _ { i }) ^ { 2} } \text { with } \bar {y}= \frac {1}{W} \sum _ { i = 1} ^ { n } w _ { i } { y _ i } $$ The R² is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer R² is to 1, the better is the model. The problem with the R² is that it does not consider the number of variables used to fit the model.
- **Adjusted R²**: The adjusted determination coefficient for the model. The adjusted R² can be negative if the R² is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by: $$ \hat { R } ^ { 2} = 1- (1- R ^ { 2}) \frac { W - 1} { W - p - 1}\cdot $$ The adjusted R² is a correction to the R² which considers the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by: $$ MSE = \frac { 1} { W - p ^ { * } } \sum _ { i = 1} ^ { x } w _ { i} (y _ { i } - \hat { y } _ { i }) ^ { 2}\cdot $$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

- **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows: $$ MAPE = \frac {100} {W} { \sum _ { i = 1} ^ { n } w _ { i } { {\lvert} \frac {y _ { i } - \hat { y } _ { i } } {y _ { i }} {\rvert} } }\cdot $$

- **DW**: The Durbin-Watson statistic is defined by: $$ D W = \frac { \sum _ { i = 2} ^ { x } [(y _ { i } - \hat { y } _ { i }) - (y _ { i - 1} - \hat { y } _ { i - 1})] ^ { 2} } { \sum _ { i = 1} ^ { x } w _ { i } (y _ { i } - \hat { y } _ { i }) ^ { 2} }\cdot $$ This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- **Cp**: Mallows Cp coefficient is defined by: $$ Cp = \frac { SSE } { \hat { \sigma } } + 2p ^ { * } - W, $$ where SSE is the sum of the squares of the errors for the model with p explanatory variables and $\hat{\sigma}$ is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to p*, the less the model is biased.

- **AIC**: Akaike's Information Criterion is defined by: $$ A I C = W \operatorname{ln} (\frac { S S E } { W }) + 2p ^ { * }\cdot $$ This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz's Bayesian Criterion is defined by: $$ S B C = W \operatorname{ln} (\frac { S S E } { W }) + \operatorname{ln} (W) p ^ { * }\cdot $$ This criterion, proposed by Schwarz (1978) is similar to the AIC and, like this, the aim is to minimize it.

- **PC**: Amemiya's Prediction Criterion is defined by: $$ P C = \frac { (1- R ^ { 2}) (W + p ^ { * }) } { W - p ^ { * } }\cdot $$ This criterion, proposed by Amemiya (1980) is used, like the adjusted R^2 to take account of the parsimony of the model.

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

If the Type I/III SS (SS: Sum of Squares) option is activated, the corresponding tables are displayed.

The table of **Type I SS** values is used to visualize the influence that progressively adding explanatory variables has on the fitting of the model, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all

the other variables already being in the model. The sums of squares in the Type I table always add up to the model SS. Note: the order in which the variables are selected in the model influences the values obtained.

The table of **Type III SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, expect those were the effect is present (interactions), as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. Type II and Type III are identical if there are no interactions or if the design is balanced.

The **parameters of the model** table displays the estimate of the parameters, the corresponding standard error, the Student's t, the corresponding probability, as well as the confidence interval.

The table of **standardized coefficients** (also called beta coefficients) is used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the residuals, the confidence. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the random values being larger. If the validation data have been selected, they are displayed at the end of the table.

The **charts** which follow show the results mentioned above. If there is only one explanatory variable in the model, the first chart displayed shows the observed values, the regression line and both types of confidence interval around the predictions. The second chart shows the standardized residuals as a function of the explanatory variable. In principle, the residuals should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a problem with the model.

The three charts displayed next show respectively the evolution of the standardized residuals as a function of the dependent variable, the distance between the predictions and the observations (for an ideal model, the points would all be on the bisector), and the standardized residuals on a bar chart. The last chart quickly shows if an abnormal number of values are outside the interval]-2, 2[given that the latter, assuming that the sample is normally distributed, should contain about 95% of the data.

If multiple comparison tests have been requested, the corresponding results are then displayed.

Finally, tables associated to the repeated measures analysis are displayed:

Mauchly's sphericity test can be used to test the sphericity of the covariance matrix between repetitions. It has a small power and should not be trusted with small samples. In this table, Greenhouse-Geisser and Huynt-Feldt epsilons can also be found. The closer they are to one, the more spherical the covariance matrix is.

The test of within-subject effects is then displayed. It shows which factor has a significant effect across repetition.

The test of between-subject effects is then displayed. It shows which factor has an effect which is significantly different from one subject to another and not from one repetition to another.

Example

A tutorial on repeated measures ANOVA is available on the XLSTAT Help Center:

http://www.xlstat.com/demo-anorep2.htm

References

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.

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. & Robotten M. (1997). Analysis of Variance for Sensory Data, John Wiley & 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.

Mixed Models

Use this tool to build ANOVA models with repeated factors, random components or repeated measures.

In this section:

Description

Mixed models are complex models based on the same principle as general linear models. They make it possible to take into account, on the one hand, the concept of repeated measurement and, on the other hand, that of random factor. The explanatory variables could be as well quantitative as qualitative. Within the framework of the mixed models, the explanatory variables are often called factors. XLSTAT uses mixed models to carry out repeated measures ANOVA.

A mixed model is written as follows:

$$y = X\beta + Z\gamma + \epsilon \quad (1)$$

where y is the dependent variable, X gathers all fixed effects (these factors are the classical OLS regression variables or the ANOVA factors), β is a vector of parameters associated with the fixed factors, Z is a matrix gathering all the random effects (factors that cannot be set as fixed), γ is a vector of parameters associated with the random effects and ϵ is an error vector. The main difference between general linear model and mixed model is that$\gamma \sim N(0, G(\theta_G))$ et $\epsilon \sim N(0, R(\theta_R))$.

We have:

$$E\begin{bmatrix}\gamma \\ \epsilon\end{bmatrix} = \begin{bmatrix}0 \\ 0\end{bmatrix} \text{ and } Var\begin{bmatrix}\gamma \\ \epsilon\end{bmatrix} = \begin{bmatrix}G & 0 \\ 0 & R\end{bmatrix}$$

The variance of y is written as follows: $Var(y) = V(\theta) = Z'GZ + R$, where θ is a vector of parameters associated with the unknown parameters of G and R. We have $y \sim N(X\beta, V(\theta))$.

According to the model to be estimated, the matrices R and G will have different forms:

- For a classical linear model, we have: $Z = 0$ and $R = \sigma^2 I_n$.
- For a repeated measures ANOVA, we have: $Z = 0$ and $cov(\epsilon) = R(\theta)$, where R is a square-block matrix with a user-defined design. Each block gathers the covariance between different measures on the same subject (which are correlated). Explanatory variables are all qualitative.
- For a random component model, we have $cov(\gamma) = G$, where G is a matrix with a user-defined design.

The following table shows the designs implemented in XLSTAT for the R and G matrices (dimension $p \times p$):

Structure de covariance	Nombre de paramètres	Formule
Variance components	Number of random factors (if no random factor =1)	$\sigma_{ij} = \sigma_k^2 I_{(i=j)}$, k is the random effect associated with the row 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)	$\sigma_{ij} = \sigma_{\lvert i-j \rvert+1} I_{(\lvert i-j \rvert<q)}$

Parameters estimation is performed by using the maximum likelihood approach. There exist two methods: the classical maximum likelihood (ML) and the restricted maximum likelihood (REML). The latter is the default in XLSTAT. The likelihood function is given by:

$$l_{REML}(G,R) = -\frac{1}{2} log \mid V \mid -\frac{1}{2} log \mid X'V^{-1}X \mid -\frac{1}{2} r'V^{-1}r - \frac{n-p}{2} log(2\pi) \quad (2)$$

where $r = y - X\hat{\beta}$. The parameters are obtained by using the first and second derivatives of $l_{REML}(G,R)$. For the details of these matrices, one can see Wolfinger, Tobias and Sall (1994). The use of an analytical method to obtain the θ parameters is not possible. XLSTAT does not profile the variance during the computation and initial values of the covariance matrix are the variances obtained with the general linear model. We thus use the iterative Newton-Raphson algorithm in order to obtain an estimate of θ. Once θ is obtained, the coefficients β and γ are calculated by solving the following equation system:

$$\begin{bmatrix} X'\hat{R}^{-1}X & X'\hat{R}^{-1}Z \\ Z'\hat{R}^{-1}X & Z'\hat{R}^{-1}Z + \hat{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X'\hat{R}^{-1}y \\ Z'\hat{R}^{-1}y \end{bmatrix} \quad (3)$$

We obtain:

$$\begin{aligned}\hat{\beta} &= \left(X'\hat{V}^{-1}X\right)^{-} X'\hat{V}^{-1}y \\ \hat{\gamma} &= \hat{G}Z'\hat{V}^{-1}\left(y - X\hat{\beta}\right)\end{aligned} \tag{4}$$

where $()^{-}$ is the generalized inverse of the matrix. The interpretation of the model is made in the same way than in the linear case.

Data format

Within the framework of mixed models, the data will have a specific format:

If there are no repeated measurements, then there will be one column per variable associated to each fixed effect and one column per variable associated to each random effect.

If there are repeated measurements, all the repetitions will have to be one after the other. We cannot have one column for each repetition. We thus define a factor identifying each repetition and another factor identifying the subject to be treated in each repetition. Thus, for a data set with 3 repetitions and 2 subjects and an explanatory variable X measured at times T_1 and T_2 for the two subjects, we have the following table:

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 makes it possible to select a repeated factor and a subject factor. These factors must be qualitative and are necessary for repeated measures ANOVA, and available for mixed models.

Interactions

By interaction is meant an artificial factor (not measured) that reflects the interaction between at least two measured factors. For example, if we carry out treatment on a plant, and tests are carried out under two different light intensities, we will be able to include in the model an interaction factor treatment*light which will be used to identify a possible interaction between the two factors. If there is an interaction between the two factors, we will observe, for example, a significantly higher effect on the plants when the light is strong and the treatment is of type 2

while the effect is average for low light and treatment 2 or strong light and treatment 1 combinations.

To make a parallel with linear regression, the interactions are equivalent to the products between the continuous explanatory values. For qualitative variables it is a little more complex, and constraints must be defined to avoid multicolinearities in the model (see below). However, the notation used to represent the interaction between factor A and factor B is A*B.

The interactions to be used in the model can be easily defined in XLSTAT.

Constraints

During the calculations, each factor is broken down into a sub-matrix containing as many columns as there are categories in the factor. Typically, this is a full disjunctive table. Nevertheless, the breakdown poses a problem: if there are g categories, the rank of this sub-matrix is not g but $g-1$. This leads to the requirement to delete one of the columns of the sub-matrix and possibly to transform the other columns. Several strategies are available depending on the interpretation we want to make afterwards:

1) **a1=0**: the parameter for the first category is null. This choice allows us force the effect of the first category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group 1.

2) **an=0**: the parameter for the last category is null. This choice allows us force the effect of the last category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group g.

Note: even if the choice of constraint influences the values of the parameters, it has no effect on the predicted values and on the different fitting statistics.

Inference and tests

XLSTAT allows computing the type I, II and III tests of the fixed effects. The principle of these tests is the same one as in the case of the linear model. Nevertheless, their calculation differs slightly. All these tests are based on the following F statistics:

$$F = \frac{\hat{\beta}' L' (L(X'\hat{V}^{-1}X)^{-}L')^{-}L\hat{\beta}}{r}$$

where L is a specific matrix associated with each fixed effect and it differs depending on the type of test. We have $r = rang(L(X'\hat{V}^{-1}X)^{-}L')$. A p-value is obtained using the Fisher distribution with $Num.DF$ and $Den.DF$ degrees of freedom. We have $Num.DF = rank(L)$ and $Den.DF$ depends on the estimated model. XLSTAT uses:

- The *contain* method if a random effect is selected, we have: $Den.DDL = N - rang(XZ)$.

- The *residual* method if no random effect is selected, we have: $\text{Den}.DDL = n - \text{rang}(X)$.
- The *Satterthwaite's approximation* method in case of unbalanced dataset : $\text{Den}.DDL = \frac{2E}{E-q}$.

$$\text{with } E = \sum_{m=1}^{q} \frac{v_m}{v_m - 2} I(v_m > 2); \quad v_m = \frac{2(D_m)^2}{g'_m A g_m};$$

where D_m is the m^{th} diagonal element of D and g_m is the gradient of $l_m C g l'_m$ with respect to θ, evaluated at $\hat{\theta}$.

D_m is from the spectral decomposition of $L\hat{C}L' = P'DP$, where P is an orthogonal matrix of eigenvectors and D is a diagonal matrix of eigenvalues, both of dimension $q \times q$. l_m is the m^{th} row of PL.

$C = (X'V^{-1}X)^-$, where $()^-$ is the generalized inverse of the matrix and C corresponds to the variance/covariance matrix of fixed parameters and θ is a vector of unknow variance parameters (random and residual parts of mixed model).

Multiple Comparisons Tests (only for repeated measures ANOVA)

As in classical ANOVA, in repeated measures ANOVA multiple comparisons can be performed. It is aimed at checking whether the various categories of a factor differ significantly or not. For example, in the case where four treatments are applied to plants, we want to know not only if the treatments have a significant effect, but also if the treatments have different effects.

Numerous tests have been proposed for comparing the means of categories. The majority of these tests assume that the sample is normally distributed. XLSTAT provides the main tests. In the case of repeated measures ANOVA, standard deviations are obtained using the maximum likelihood estimates. For more details on the tests, please see the description section of the ANOVA help.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Dependent variables:

Quantitative: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

X / Explanatory variables:

Quantitative: Activate this option to perform an ANCOVA analysis. Then select the quantitative explanatory variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Select the qualitative explanatory variables (the factors) in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

Regression weights: Activate this option if you want to include weights in the model's equation. If you do not activate this option, the weights will be considered as 1. Weights must be greater

than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Fixed constant: Activate this option to fix the constant of the regression model to a value you then enter (0 by default).

Tolerance: Activate this option to prevent the OLS regression calculation algorithm taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

Constraints: Details on the various options are available in the description section.

a1 = 0: Choose this option so that the parameter of the first category of each factor is set to 0.

an = 0: Choose this option so that the parameter of the last category of each factor is set to 0.

Repeated measures: Activate this option if you want to include a repeated factor in your model.

Covariance structure: Choose the covariance structure you want to use for the R matrix. XLSTAT offers: Autoregressive(1), Compound Symmetry, Toeplitz, Toeplitz(q), Unstructured and Variance Components. Details on the various options are available in the description section.

Random effect: (only with mixed models): Activate this option if you want to include a random effect in your model.

Covariance structure: Choose the covariance structure you want to use for the R matrix. XLSTAT offers: Autoregressive(1), Compound Symmetry, Toeplitz, Toeplitz(q), Unstructured and Variance Components. Details on the various options are available in the description section.

Satterthwaite's t-tests: Activate this option if you want to compute t-tests for fixed coefficients *Beta* using the Satterthwaite formula for denominator degrees of freedom. F-tests concerning tests of fixed effect will be computed according to this method as well. Mixed models with unbalanced datasets are automatically computed using the Satterthwaite approximation.

Newton-Raphson's starting values: Choose the method to set initial values to the iterative Newton-Raphson algorithm. Either use the OLS option to specify ordinary least squares starting values or MIVQUE0 (Minimum Variance Quadratic Unbiased Estimation). When covariance structure is basic, MIVQUE0 estimates are the REML estimates. For more complex covariance structure, MIVQUE0 could be choosen in order to be as close as possible to the population values so that the optimization routine can converge at reasonable estimates.

Estimation method: choose between REML and ML to estimate your model. Details are available in the description section.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

General:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the correlation matrix for quantitative variables (dependent or explanatory).

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Covariance parameters: Activate this option to display the table of the covariance parameters.

Null model likelihood ratio test: Activate this option to display the results of the null model likelihood ratio test.

Fixed effects coefficients: Activate this option to display the table of the fixed effects coefficients.

Random effects coefficients (only with mixed models)**:** Activate this option to display the table of the random effects coefficients.

Type III tests of fixed effect: Activate this option to display the results of the type III tests of the fixed effect.

Type I tests of fixed effect: Activate this option to display the results of the type I tests of the fixed effect.

Type II tests of fixed effect: Activate this option to display the results of the type II tests of the fixed effect.

R matrix: Activate this option to display the error covariance matrix R for the first subject.

G matrix (only with mixed models)**:** Activate this option to display the random effects covariance matrix G.

Residuals:

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

- **Raw residuals**: Activate this option to display the raw residuals in the predictions and residuals table.
- **Studentized residuals**: Activate this option to display the studentized residuals in the predictions and residuals table.
- **Pearson residuals**: Activate this option to display the Pearson residuals in the predictions and residuals table.

Comparisons (only for repeated measures ANOVA):

Multiple comparisons:

Information on the multiple comparison tests is available in the description section.

Apply to all factors: Activate this option to compute the selected tests for all factors.

Use least squares means: Activate this option to compare the means using their least squares estimators (obtained from the parameters of the model). If this option is not activated, the means are computed using their estimation based on the data.

Sort up: Activate this option to sort the compared categories in increasing order, the sort criterion being their respective means. If this option is not activated, the sort is decreasing.

Pairwise comparisons: Activate this option then choose the comparison methods.

Factors and interactions dialog box

Once the first dialog box disappears, a second one appears, to allow you to what type of factor each factor corresponds. The layout and the aim of the dialog box depends the type of ANOVA you want to run:

If repeated measures were selected, it is necessary to select the fixed factors (fixed effects), a repeated factor, and a subject factor.

If random effects have been selected, it is necessary to specify which factors are fixed and which are random.

If both repeated measures and random effects have been selected, it is necessary to specify which factors are fixed, which are random, and to define which is the repeated factor and which is subject factor.

Each factor must be selected only once. Repeated and subject factors must be qualitative.

Results

Summary statistics: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the dependent variables (in blue) and the quantitative explanatory variables. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

Correlation matrix: This table is displayed to give you a view of the correlations between the various variables selected.

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.
- **AIC**: the Akaike's Information Criterion (AIC) is defined by:

$$AIC = -2l(\theta) + 2d$$

where l is the likelihood function and d equals the number of parameters to be estimated. This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **AICC**: This criterion derived from the AIC is defined by:

$$AICC = -2l(\theta) + 2dn/(n-d-1)$$

- **SBC**: Schwarz's Bayesian Criterion is defined by:

$$SBC = -2l(\theta) + d\ln(n)$$

This criterion, proposed by Schwarz (1978) is similar to the AIC and the aim is to minimize it.

- **CAIC**: This criterion (Bodzogan, 1987) is defined by:

$$CAIC = -2l(\theta) + d(\ln(n) + 1)$$

- **Iterations**: This value gives the number of iteration necessary to reach the convergence of the Newton-Raphson algorithm.
- **Covariance parameters**: This value gives the number of parameters to be estimated in the covariance matrix V.

- **Number of fixed effects:** This value gives the number of selected fixed effects.
- **Number of random effects:** This value gives the number of selected random effects.

Covariance parameters – Repeated factors: This table displays the covariance parameters associated to the repeated factor. For each parameter, the corresponding standard error, the Z statistic, the corresponding probability, as well as the confidence interval are presented.

Covariance parameters – Random factors (only with mixed models)**:** This table displays covariance parameters associated to the random. For each parameter, the corresponding standard error, the Z statistic, the corresponding probability, as well as the confidence interval are presented.

The **null model likelihood ratio test table** compares the likelihood of the null model and the likelihood of the selected model. The likelihood ration, the Chi-square statistic and the corresponding probability are displayed.

The **model parameters** table displays the estimate of the parameters, the corresponding standard error, the Student’s t, the corresponding probability, as well as the confidence interval

The **random effects coefficients** (only with mixed models) table displays the estimate of the random effects parameters, the corresponding standard error, the number of degrees of freedom, the Student’s t, the corresponding probability and confidence interval

If the Type I tests and Type III tests of fixed effects have been requested, the corresponding tables are displayed.

The table of **Type I tests of fixed effects** values is used to evaluate the influence of sequentially adding explanatory variables on the fit of the model, through the Fisher's F or its corresponding p-value. The lower the probability, the larger the contribution of the variable to the model (given that all the previously added variables are in the model).

Note: the order in which the variables are selected in the model influences the values obtained.

The table of **Type III tests of fixed effects** values is used to evaluate the impact of removing an explanatory variable, all other variables being retained, in terms of Fisher's F and its corresponding p-value. The lower the probability, the larger the contribution of the variable to the model, all other variables already being in the model.

Note: unlike Type I tests of fixed effects, the order in which the variables are selected in the model does hot have any influence on the values obtained.

The **predictions and residuals** table shows, for each observation, its weight, the observed value of the dependent variable, the model's prediction, the residuals, the confidence intervals. Several types of residuals are displayed:

- Raw residuals:

$$r_i = y_i - x_i' \hat{\beta}$$

- Studentized residuals:

$$r_i^{stud} = \frac{r_i}{\sqrt{var(r_i)}}$$

- Pearson's residuals:

$$r_i^{stud} = \frac{r_i}{\sqrt{var(y_i)}}$$

If one or more random effects are selected, we have:

- Conditional raw residuals:

$$r_i^{cond} = r_i - z_i'\hat{\gamma}$$

- Studentized conditional residuals:

$$r_i^{cond/stud} = \frac{r_i^{cond}}{\sqrt{var(r_i^{cond})}}$$

Pearson's conditional residuals:

$$r_i^{cond/pearson} = \frac{r_i}{\sqrt{var(y_i)}}$$

If multiple comparison tests have been requested, the corresponding results are then displayed.

Example

A tutorial on repeated measures ANOVA is available on the XLSTAT Help Center:

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

A tutorial on random component model is available on the XLSTAT Help Center:

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

References

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

Use this model to carry out a MANOVA (Multivariate ANalysis Of VAriance) of two or more balanced or unbalanced factors. The advanced options enable you to choose the confidence level and to take into account interactions between the factors. Multivariate tests can be calculated.

In this section:

Description

The MANOVA uses the same conceptual framework as the ANOVA. The main difference comes from the nature of the dependent variables: instead of one, we can study many of them. With the MANOVA, explanatory variables are often called factors. Effects of factors are estimated on a combination of several response variables.

The advantage of the MANOVA as opposed to several simultaneous ANOVAs lies in the fact that it takes into account correlations between response variables, which results in a richer use of the information contained in the data.

The MANOVA tests the presence of significant differences among combinations of levels of factors on several response variables. MANOVA also enables the simultaneous tests of all hypotheses tested by an ANOVA and is more likely to detect differences between levels of factors.

Furthermore, the computation of several ANOVAs instead of one MANOVA increases the Type I error, which is the probability that the null hypothesis will be wrongly rejected.

The potential covariation between response variables is not taken into account with several ANOVAs. Instead, the MANOVA is sensitive to both the difference of averages between levels of factors and the covariation between explanatory variables. And a potential correlation between response variables is more likely to be detected when these variables are studied together, as is the case with a MANOVA.

Let's consider as an illustrative example a two-way MANOVA. The model is written as followed:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk} \quad (1)$$

where y_{ijk} is the kth observation of the ith level of A and jth level of B, ϵ is the error of the model.

The hypotheses used in a MANOVA are identical to those used in linear regression: errors ϵ follow the same normal distribution $N(0, s)$ and are independent.

To use the various tests proposed in the results of linear regression, one should check retrospectively that the underlying hypotheses have been correctly verified. The normality of the residuals can be checked by analyzing certain charts or by using a normality test. The independence of the residuals can be checked by analyzing certain charts or by using the Durbin Watson test.

Interactions

An “interaction” is an artificial factor (not measured) that reflects the interaction between at least two measured factors. For example, if we carry out treatment on a plant, and tests are carried out under two different light intensities, we will be able to include in the model an interaction factor treatment*light which will be used to identify a possible interaction between the two factors. If there is an interaction between the two factors, we will observe a significantly larger effect on the plants when the light is strong and the treatment is of type 2, while the effect is average for weak light, treatment 2 and strong light, treatment 1 combinations.

To make a parallel with linear regression, the interactions are equivalent to the products between the continuous explanatory values, although here obtaining interactions requires nothing more than simple multiplication between two variables. However, the notation used to represent the interaction between factor A and factor B is A*B.

The interactions to be used in the model can be easily defined in XLSTAT.

Balanced and unbalanced MANOVA

We talk of balanced MANOVA when the numbers of categories are equal for all combinations of factors. When the numbers of all categories for one of the combinations of factors are not equal, then the MANOVA is said to be unbalanced. XLSTAT can handle both cases.

Constraints

During the calculations, each factor is broken down into a sub-matrix containing as many columns as there are categories in the factor. Typically, this is a full disjunctive table. Nevertheless, the breakdown poses a problem: if there are g categories, the rank of this sub-matrix is not g but $g - 1$. This leads to the mandatory deletion of one of the columns of the sub-matrix and possibly the transformation the other columns. The strategy taken in XLSTAT is the following:

a1=0: the parameter for the first category is null. This choice allows us to force the effect of the first category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group 1.

Moreover, the number of observations should be equal to at least the sum of the number of dependent variables and the number of factors and interactions included in the model (+1).

Multivariate tests

One of the main application of the MANOVA is multivariate comparison testing where parameters for the various categories of a factor are tested to be significantly different or not. For example, in the case where four treatments are applied to plants, we want to know if treatments have a significant effect and also if treatments have different effects.

Numerous tests have been proposed to compare means of each category. Most of them rely on the relationships that exist between the error matrix E and the matrix symbolizing the tested hypotheses H that is the eigenvalues of the matrix $E^{-1}H$. XLSTAT provides the main tests including:

Wilks Lambda test: the likelihood ratio test statistic also known as Wilks Lambda (1932) is given by:

$$Lambda = \prod_{i=1}^{m} \frac{1}{1+\lambda_i}$$

The null hypothesis is rejected for small values of Lambda, indicating that the error E is small compared to the total SSCP matrix $E+H$. This test is the most frequently used.

When there are two levels, the test is equivalent to the Fisher test mentioned previously. If the number of levels is less than or equal to three, the test is exact. The Rao approximation is required from four classes to obtain a statistic approximately distributed according to a Fisher distribution.

Hotelling-Lawley's Trace test:

$$T_{HL} = \sum_{i=1}^{m} \lambda_i = Trace(E^{-1}H)$$

A large H compared to E indicates a larger trace. Hence, the null hypothesis of no effects i0073 is rejected for large values of T_{HL}. This test is efficient if all factors have exactly two levels.

Pillai's Trace test:

$$T_P = \sum_{i=1}^{m} \frac{\lambda_i}{1+\lambda_i} = \mathrm{Trace}((E+H)^{-1}H)$$

As with Hotelling-Lawley's trace, the null hypothesis is rejected for large values of T_p, indicating a large H relative to E. This test is efficient if all samples have the same number of observations.

Roy's greatest root test:

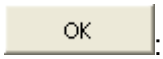
$$\lambda_{max} = \max_{1 \leq i \leq m} \lambda_i$$

The computed p-value for this test is always smaller than from other tests. Roy's test is a powerful but not robust test. For this reason, it's not recommended.

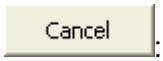
The last 3 tests detailed above (Hotelling-Lawley's Trace test, Pillai's Trace test, Roy's greatest root test) are less used than the Wilks' Lambda test and are based on the Fisher distribution for calculating p-values.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Dependent variables:

Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

X / Explanatory variables:

Select the qualitative explanatory variables (the factors) in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as is. 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

Options tab:

Confidence interval (%): Enter the level of significance for the different calculated tests.

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

Do not allow missing values observations: Activate this option to avoid missing values.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Means by factor level: Activate this option to display the means by factor level.

SSCP matrices: Activate this option to display the SSCP matrices for factors and interactions.

Eigen values: Activate this option to display the eigenvalues of SSCP matrices get for each factor and interaction.

Tests results : Activate this option to display the results of statistical tests.

- **Wilks lambda test (Rao approximation)** : Activate this option to display the results of the Lambda statistic and the associated p-value.

- **Hotelling-Lawley trace** : Activate this option to display the results of the Hotelling-Lawley trace test.
- **Pillai's trace** : Activate this option to display the results of the Pillai trace test.
- **Largest Roy Root** : Activate this option to display the results of the largest Roy root test.

Graphs tab:

Means charts: Activate this option to display the means of level factor with histograms.

Results

Summary statistics: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the dependent variables (in blue) and the quantitative explanatory variables. For qualitative explanatory variables, the names of the various categories are displayed together with their respective frequencies.

Means of level factor: This table provides for each level factor and quantitative variable the mean value.

SSCP matrices: These tables are displayed to give a general view of the effects of the factors and interactions between factors.

Wilks' test (Rao's approximation): This table provides the results of Wilks' Lambda test which tests the hypothesis of equality of the mean vectors for the different levels. The details of this test are given in paragraph Description.

Hotelling-Lawley test: This table provides the results of Hotelling-Lawley trace test, which tests the hypothesis of equality of the mean vectors for the different levels. The details of this test are given in paragraph Description.

Pillai's test: This table provides the results of Pillai's trace test, which tests the hypothesis of equality of the mean vectors for the the different levels. The details of this test are given in paragraph Description.

Roy's test: This table provides the results of Roy's greatest root test which tests the hypothesis of equality of the mean vectors for the different levels. The details of this test are given in paragraph Description.

Example

A tutorial on one-way MANOVA is available on the XLSTAT Help Center:

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

References

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.

Logistic regression

Use logistic regression to model a binomial, multinomial or ordinal variable using quantitative and/or qualitative explanatory variables.

In this section:

Description

Logistic regression is a frequently used method because it allows to model binomial (typically binary) variables, multinomial variables (qualitative variables with more than two categories) or ordinal (qualitative variables whose categories can be ordered). It is widely used in the medical field, in sociology, in epidemiology, in quantitative marketing (purchase or not of products or services following an action) and in finance for risk modeling (scoring).

The principle of the logistic regression model is to explain the occurrence or not of an event (the dependent variable noted Y) by the level of explanatory variables (noted X). For example, in the medical field, we seek to assess from what dose of a drug, a patient will be cured.

Binomial logistic regression

Logistic and linear regression belong to the same family of models called GLM (*Generalized Linear Model*): in both cases, an event is linked to a linear combination of explanatory variables.

For linear regression, the dependent variable follows a normal distribution $N(\mu, \sigma)$ where μ is a linear function of the explanatory variables. For logistic regression, the dependent variable, also called the response variable, follows a Bernoulli distribution of parameter p (p is the mean probability that an event will occur) when the experiment is repeated once, or a $Binomial(n, p)$ distribution if the experiment is repeated n times (for example the same dose given to n patients). The probability parameter p is here a function of a linear combination of explanatory variables.

The most common functions used to link probability p to the explanatory variables are the logistic function (we refer to the *Logit* model) and the standard normal distribution function (the *Probit* model). Both these functions are perfectly symmetric and sigmoid: XLSTAT provides two other functions: the complementary Log-log function which is closer to the upper asymptote, and the Gompertz function which, on the contrary, is closer the axis of abscissa.

The analytical expression of the models is as follows:

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. Complementary Log-log: $p = 1 - \exp[-\exp(\beta X)]$
4. Gompertz: $p = exp[-exp(-\beta X)]$

where βX represents the linear combination of the q explanatory variables (including constants).

The knowledge of the distribution of the event being studied allows to compute the likelihood of the sample. To estimate the βX parameters of the model (the coefficients of the linear function), we need to maximize the likelihood function. Contrary to linear regression, an exact analytical solution does not exist. So an iterative algorithm has to be used. XLSTAT uses a Newton-Raphson algorithm. The user can set the maximum number of iterations, the convergence threshold, or the maximum time spent if desired.

In the case of binomial logistic regression, when we know the probability p associated with a category (typically 1), we deduce the probability $1 - p$ of the other category (typically 0). The reference (or control) category is by default the first category in the alphabetical order, that is to say 0 if one is in a (0/1) case. The model is calculated for the category which is not the reference one, but it is straightforward to obtain the probability for the reference category.

In most software, the calculation of **confidence intervals** for the model parameters is as for linear regression assuming that the parameters are normally distributed. XLSTAT also offers the alternative "*Likelihood ratio*" method (Venzon and Moolgavkar, 1988). This method is more reliable as it does not require the assumption that the parameters are normally distributed. Being iterative, however, it can slow down the calculations.

Multinomial logistic regression

The principle of multinomial logistic regression is to explain or predict a variable that can take J alternative values (the J categories of the variable), as a function of explanatory variables. The binomial case seen previously is therefore a special case where $J = 2$.

Within the framework of the multinomial model, a control category must be selected. Ideally, we will choose what corresponds to the "basic" or "classic" or "normal" situation. The estimated coefficients will be interpreted according to this control category. For ease of writing, the equations below are written considering the first category as the reference category.

The model proposed by XLSTAT to relate the probability of occurrence of an event to the explanatory variables is the logit model which is one of the four models proposed for the binomial case. The analytical expression of the model for categories 2 to J is given below:

$$\log(\frac{p(Y=j|x_i)}{p(Y=1|x_i)}) = \alpha_j + \beta_j X_i \ , \quad i = 2...j$$

$$p(Y=j|X_i) = \frac{exp(\alpha_j + \beta_j X_i))}{1+\sum_{k=2}^{J}\exp(\alpha_k+\beta_k X_i)}, \quad i=2...j$$

For the 1st category, we have:

$$p(Y=1|x_i) = \frac{1}{1+\sum_{k-2}^{J}\exp(\alpha_k+\beta_k X_i)}$$

The likelihood of the sample is given by:

$$l(\alpha,\beta) = \sum_{i=1}^{n}\sum_{j=1}^{J} y_{ij}\log(p(Y=j|x_i))$$

To estimate the α and β parameters of the model (the coefficients of the linear function), we try to maximize the likelihood function. Contrary to linear regression, an exact analytical solution does not exist. XLSTAT uses the Newton-Raphson algorithm to iteratively find a solution.

Ordinal logistic regression

The principle of ordinal logistic regression is to explain or predict a variable that can take J ordered alternative values (only the order matters, not the differences), as a function of a linear combination of the explanatory variables. Binomial logistic regression is a special case of ordinal logistic regression, corresponding to the case where $J=2$.

XLSTAT makes it possible to use two alternative models to calculate the probabilities of assignment to the categories given the explanatory variables: the logit model and the probit model. For the case of the logit model, we have:

$$\log(\frac{p(Y\leq j|x_i)}{p(Y>j|x_i)}) = \alpha_j + \beta X_i$$

We can see that there is an intercept for each category of the dependent variable Y, but unlike the multinomial model, we have only one set of β of coefficients, whatever the category of Y. The reference category is always the lowest. The probability of choosing the category j or a category lower than j is given by:

$$p(Y\leq j|x_i) = \frac{\exp(\alpha_j+\beta X_i)}{1+\exp(\alpha_j+\beta X_i)}, \quad i=1...J-1$$

For $j=J$, this probability is 1.

The probability for category j is given by:

$$p(Y=j|x_i) = p(Y\leq j|x_i) - p(Y\leq j-1|x_i)$$

As a consequence, the likelihood of the sample writes:

$$l(\alpha,\beta) = \sum_{i=1}^{n}\sum_{j=1}^{J} y_{ij}\log(p(Y \leq j|x_i) - p(Y \leq j-1|x_i))$$

To estimate the q parameters β and the $(J-1)$ parameters α_j of the model (the coefficients of the linear combination), we maximize the likelihood function. Unlike linear regression, an exact analytical solution does not exist. It is therefore necessary to use an iterative algorithm. XLSTAT uses a Newton-Raphson algorithm.

Predictions and influence diagnostics

Whatever the type of response variable, binary, sum(binary, multinomial or ordinal, XLSTAT provides for each observation the probability of assignment to each of the possible categories.

In the binomial case (binary or sum(binary)), for a given cutoff C, typically 0.5, if the probability for observation i is less than this threshold value, the observation is considered as being assigned to class 0, otherwise, it is assigned to class 1. If a cutoff other than 0.5 has been chosen, the probability of the category which is not the control one (in the binary case, is compared to the threshold value and if it is higher, then category 1 is predicted).

The notion of cutoff value does not apply in the ordinal or multinomial case and the observations are always assigned to the category for which the probability is the greatest.

If the **significance analysis** option has been activated, XLSTAT computes whether the probability of the elected category is significantly higher than the ones of the other categories. Indeed, for the decision maker, it is important to know to what extent the choice is marred by uncertainty. This practice is unfortunately too unusual and it is to encourage and facilitate it that XLSTAT displays this result.

In the binomial case, this analysis is automatically deduced from the confidence intervals of the probabilities. If the confidence interval around a probability does not include the cuttoff point, then the risk of error is limited to 5% (the percentage depends on the choice made for the size of the confidence intervals). For multinomial and ordinal cases, the calculations are more complicated and XLSTAT is the only software to perform them. The comparisons are made two by two.

The significance analysis is displayed in two columns. The first indicates whether, in the case where the predicted category is not that observed, if this change is significant or not. The second column indicates, whatever the chosen category, whether or not the probability for this category is greater than those of the other categories.

Finally, in the case of binomial dependent variables, if the option **influence diagnostics** has been selected, XLSTAT displays the table giving various statistics, in particular recommended by Pregibon (1981).

Let us denote by $\pi_{i,j}$ the probability calculated by the model that for observation i we observe the j^{th} category.

For the regression on binomial variables, the calculated indices are: * Residual: in the "binary" case, the residual e_i is given by $1-\pi_{i,j}$ where j is the observed category. The closer the residual is to 0, the better the prediction. The closer it is to 1, the worse the prediction. For the

sum case (binary), the residual corresponds to the difference between the number of observed and predicted cases. * Model residual: this residual is given by: $\tilde{e_i} = e_i/[\pi_{i,j}(1-\pi_{i,j})]$. It is sometimes refered to as the logit residual. * Standardized (std) residual: $z_i = e_i/\sqrt{\pi_{i,j}(1-\pi_{i,j})}$. This residue is also referred to as the Pearson residue. The sum of the squares of these residuals gives the fit statistic for χ^2. * Deviance: the deviance d_i is used to measure whether or not an observation influences the model. The further the value is from 0, the greater the influence. The sign indicates whether or not the observed value is less than the predicted value. The sum of the squared deviations gives $-2LL$ (LL denotes the log-likelihood). * Leverage: the *leverage* h_i makes it possible to identify observations for which the observed values are not as expected by the model and they are therefore atypical compared to a majority of other observations. * Studentized residual: they are given by $d_i = d_i/\sqrt{1-h_i}$ and combine the notions of deviance and leverage. The higher they are, the more suspect the observation is with regard to the model. * Cook's distance: it is given by $c_i = z_i^2 h_i/(1-h_i)$. The higher it is, the more the observation deserves to be studied closely, in that it is atypical and was probably not recorded under the same experimental conditions as the others. * DFBeta: this index is calculated for each explanatory variable. It makes it possible to measure the impact of each observation on the coefficients of each of the explanatory variables. The sign of the DFBeta indicates in which direction the influence takes place, and the higher the value, the more the observation influences the value of the coefficient.

Well and misclassified observations, GCI index and ROC curve

XLSTAT gives the possibility of displaying the classification table (also called confusion matrix) which allows to calculate a percentage of well-classified observations.

We denote by sensitivity (*sensivity*) the proportion of positive events well classified (true positives). The specificity (*specificity*) corresponds to the proportion well-classified negative events (true negatives). If we vary the threshold probability from which we consider that an event should be considered positive, the sensitivity and specificity vary. The points curve (1-specificity, sensitivity) is the ROC curve.

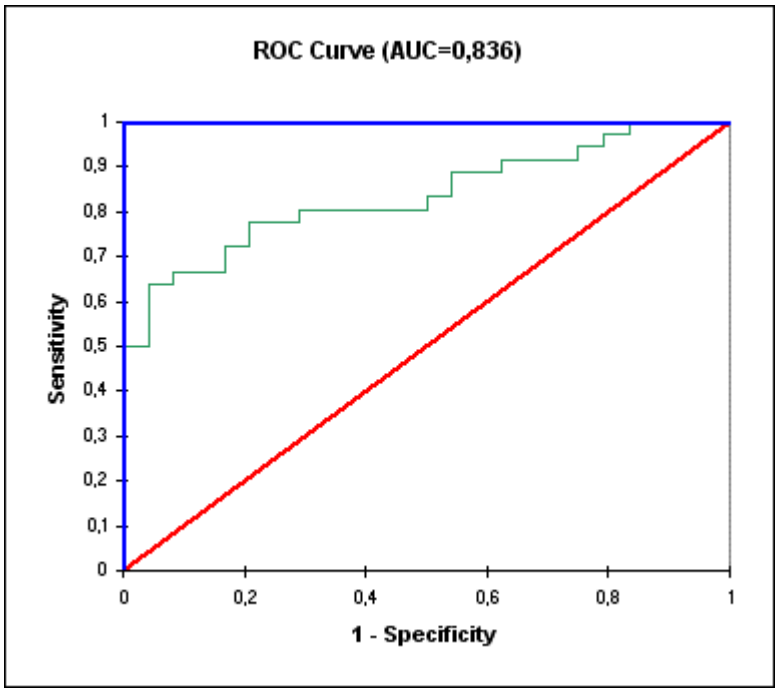
If the option **significance analysis** has been activated, a confusion matrix with an additional "Uncertain" column is added, in order to count the observations for which the predicted category is uncertain (the highest probability is not significantly different from all the others).

A table summarizes the four calculated indices. Let W be the sum of the weights of the N observations of the learning sample, BC the number of well-classified individuals, MC the number of misclassified individuals and IC the number of uncertain individuals. The table shows: *% correct = BC/W: percentage of well-classified observations (true positives) *% uncertain IC/W: percentage of observations for which the ranking is uncertain *% incorrect MC/W: percentage of misclassified observations (false positives and false negatives) * $GCI = (BC - MC + IC/2)/W$: the *Goodness of Classification Index* is derived from previous values, developed by Addinsoft in order to simply and realistically assess quality predictive of a classification model. It is expressed in %.

Special case of binary variables: Sensitivity denotes the proportion of well-classified positive events (true positives). Specificity corresponds to the proportion of well-classified negative events (true negatives). If we vary the threshold probability from which we consider that an event should be considered positive, the sensitivity and specificity vary. The curve (1-specificity,

sensitivity) is the ROC (*Receiver Operating Characteristics*) curve. It is used to visualize the performance of a model, and to compare it with that of other models. The terms used come from signal detection theory.

Example: Let us consider a binary dependent variable which indicates, for example, if a customer has responded favorably to a mail shot. In the diagram below, the blue curve corresponds to an ideal case where the n% of people responding favorably corresponds to the n% highest probabilities. The green curve corresponds to a well-discriminating model. The red curve (first bisector) corresponds to what is obtained with a random Bernoulli model with a response probability equal to that observed in the sample studied. A model close to the red curve is therefore inefficient since it is no better than a purely random model. A model below this curve would be disastrous since it would be less even than random.



The area under the curve (or *AUC*) is a synthetic index calculated for ROC curves. The AUC corresponds to the probability such that a positive event has a higher probability given to it by the model than a negative event. For an ideal model, AUC=1 and for a random model, AUC = 0.5. A model is usually considered good when the AUC value is greater than 0.7. A well-discriminating model must have an AUC of between 0.87 and 0.9. A model with an AUC greater than 0.9 is excellent.

Separation problem

In the example above, the treatment variable is used to make a clear distinction between the positive and negative cases.

	Treatment 1	**Treatment 2**
Response +	121	0
Response -	0	85

In such cases, there is an indeterminacy on one or more parameters for which the variance is as high as the convergence threshold is low which prevents a confidence interval around the parameter from being given. To resolve this problem and obtain a stable solution, Firth (1993)

proposed the use of a *penalized likelihood* function. XLSTAT offers this solution as an option and uses the results provided by Heinze (2002). If the standard deviation of one of the parameters is very high compared with the estimate of the parameter, it is recommended to restart the calculations with the "Firth" option activated.

Since version 2021.3 XLSTAT also offers the L2 penalization, which also solves this problem.

Constraints

During the calculations, each factor is broken down into a sub-matrix containing as many columns as there are categories in the factor. Typically, this is a full disjunctive table. Nevertheless, the breakdown poses a problem: if there are g categories, the rank of this sub-matrix is not g but $g-1$. This leads to the requirement to delete one of the columns of the sub-matrix and possibly to transform the other columns. Several strategies are available depending on the interpretation we want to make afterwards:

1) **a1=0**: the parameter for the first category is null. This choice allows us force the effect of the first category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group 1.

2) **an=0**: the parameter for the last category is null. This choice allows us force the effect of the last category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group g.

3) **Sum(ai)=0**: the sum of the parameters is null. This choice forces the constant of the model to be equal to the mean of the dependent variable when the ANOVA is balanced.

4) **Sum(ni.ai)=0**: the weighted sum of the parameters is null. This choice forces the constant of the model to be equal to the mean of the dependent variable even when the ANOVA is unbalanced.

Note: even if the choice of constraint influences the values of the parameters, it has no effect on the predicted values and on the different fitting statistics.

Hosmer-Lemeshow Test

The Hosmer-Lemeshow test is a goodness of fit test for a binary logit model. It uses a statistic that follows a Chi-square distribution.

The calculation of this statistic is separated into several steps:

- The sample is ordered according to the probabilities calculated from the model in a decreasing way.
- The sample is divided into k parts of equal size.
- The Hosmer-Lemeshow statistic is calculated using the following formula:

$$S_{HL} = \sum_{i=1}^{k} \frac{O(i) - n_i P(i)}{n_i P(i)(1 - P(i))}$$

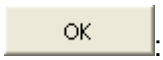
with n_i being the size of group i, $O(i)$ the number of times $Y = 1$ in group i and $P(i)$ the mean probability obtained from the model for group i.

This statistic follows a χ^2 distribution with $k - 2$ degrees of freedom. XLSTAT uses $k = 10$.

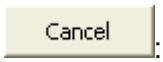
When this statistic is large and the p-value is small, then this shows a lack of fit of the model.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Response variable(s): Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

Response type: Choose the type of response variable you have selected:

- **Binary variable**: If you select this option, you must select a variable containing exactly two distinct values. If the variable has value 0 and 1, XLSTAT will see to it that the high probabilities of the model correspond to category 1 and that the low probabilities correspond to category 0. If the variable has two values other than 0 or 1 (for example Yes/No), the lower probabilities correspond to the first category and the higher probabilities to the second.
- **Sum(binary)**: If your response variable is a sum of binary variables, it must be of type numeric and contain the number of positive events (event 1) amongst those observed.

The variable corresponding to the total number of events observed for this observation (events 1 and 0 combined) must then be selected in the "Observation weights" field. This case corresponds, for example, to an experiment where a dose D (D is the explanatory variable) of a medicine is administered to 50 patients (50 is the value of the observation weights) and where it is observed that 40 get better under the effects of the dose (40 is the response variable).

- **Multinomial**: if your response variable has more than two categories, a multinomial logit model is estimated. A new field called "control category" appears. You can select the reference category.
- **Ordinal**: if your response variable has ordered categories, an ordinal logit model is estimated. The reference category is the lower category. The type of data has to be numeric with a limited number of categories.

Explanatory variables:

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If the variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Model: Choose the type of function to use (see description).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Observation weights: This field must be entered if the "sum of binary variables" option has been chosen. Otherwise, this field is not active. If a column header has been selected, check

that the "Variable labels" option has been activated.

Regression weights: Activate this option if you want to weight the influence of observations to adjust the model. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Control category: In the multinomial case, you need to choose which category is the control.

Options tab:

General sub-tab:

Algorithm:

- **Iterations**: Enter the maximum number of iterations for the Newton-Raphson algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 100.
- **Convergence**: Enter the maximum value of the evolution of the log of the likelihood from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.
- **Maximum time**: Enter the maximum that XLSTAT should spend on finding the maximum of the likelihood function. Default value: 180 seconds.
- **Penalization**: You can choose not to penalize the likelihood (default option) or to use **Firth's** penalization or the **L2** penalization (see section [description] (# description)). In the second case, please enter the **lambda** constant (Default value: 0.01).

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Tolerance: Enter the value of the tolerance threshold below which a variable will automatically be ignored.

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

LR confidence intervals: Activate this option to compute the LR confidence intervals.

Advanced sub-tab:

Model selection: Activate this option if you want to use one of the five selection methods provided:

- **Best model**: This method lets you choose the best model from amongst all the models which can handle a number of variables varying from "Min variables" to "Max Variables". Furthermore, the user can choose several "criteria" to determine the best model.

- **Criterion**: Choose the criterion from the following list: Likelihood, LR (likelihood ratio), Score, Wald, Akaike's AIC, Schwarz's SBC.

- **Min variables**: Enter the minimum number of variables to be used in the model.

- **Max variables**: Enter the maximum number of variables to be used in the model.

Note: although XLSTAT uses a very powerful algorithm to reduce the number of calculations required as much as possible, this method can require a long calculation time. The method is only available for binary logit model.

- **Stepwise (Forward)**: The selection process starts by adding the variable with the largest contribution to the model. If a second variable is such that its entry probability is greater than the **entry threshold value**, then it is added to the model. After the third variable is added, the impact of removing each variable present in the model after it has been added is evaluated. If the probability of the calculated statistic is greater than the **removal threshold value**, the variable is removed from the model.

- **Stepwise (Backward)**: This method is similar to the previous one but starts from a complete model.

- **Forward**: The procedure is the same as for stepwise selection except that variables are only added and never removed.

- **Backward**: The procedure starts by simultaneously adding all variables. The variables are then removed from the model following the procedure used for stepwise selection.

Classes weight correction: If the number of observations for the various classes for the dependent variables are not uniform, there is a risk of penalizing classes with a low number of observations in establishing the model. To get over this problem, XLSTAT has two options:

- **Automatic**: Correction is automatic. Artificial weights are assigned to the observations in order to obtain classes with an identical sum of weights.

- **Corrective weights**: You can select the weights to be assigned to each observation.

Constraints: Details on the various options are available in the description section.

- **a1 = 0**: Choose this option so that the parameter of the first category of each factor is set to 0.

- **an = 0**: Choose this option so that the parameter of the last category of each factor is set to 0.

- **Sum (ai) = 0**: for each factor, the sum of the parameters associated with the various categories is set to 0.

- **Sum (ni.ai) = 0**: for each factor, the sum of the parameters associated with the various categories weighted by their frequencies is set to 0.

Comparisons: Activate this option if you want to compare the parameters corresponding to the categories of the qualitative explantory variables.

Validation tab:

Validation: Activate this option if you want to use a sub-sample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

Quantitative: Activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

Qualitative: Activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

Observations labels: Activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Variable labels: Activate this option if the first row of the data selections includes a header.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT stops and prompts you if missing data are detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the explanatory variables correlation matrix.

Multicolinearity statistics: Activate this option to display the multicolinearity statistics for all explanatory variables.

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Type II analysis: Activate this option to display the type II analysis of variance table.

Hosmer-Lemeshow test: Activate this option to display the results of the Hosmer-Lemeshow test.

Model coefficients: Activate this option to display the table of coefficients for the model. Optionally, **confidence intervals** of type "*profile likelihood* " can be calculated (see description).

Equation: Activate this option to display the equation for the model explicitly.

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Covariance matrix: Activate this option to display the covariance matrix of the estimated parameters of the model.

Marginal effects: Activate this option if you want the marginal effects at the means to be displayed. These effects make it possible to measure the impact of each explanatory variable when all the others are set at their mean.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

- **Independent model**: Activate this option to display the results corresponding to the basic model (called independent) where the predicted probability simply corresponds to the observed frequency of each category of the dependent variable.
- **Confidence intervals**: activate this option to display the confidence intervals. These are only displayed in cases where the response variable is binomial.
- **Significance analysis**: activate this option to analyze whether the probability associated with the predicted modality is significantly different from that calculated for other modalities (see the description section).

Classification table: Activate this option to display the posterior observation classification table using a **cutoff point** to be defined (default value 0.5).

Probability analysis: If only one explanatory variable has been selected, activate this option so that XLSTAT calculates the value of the explanatory variable corresponding to various probability levels.

Multiple comparisons: This option is only active if qualitative explanatory variables have been selected. Activate this option to display the results of the comparison tests.

Charts tab:

Regression charts: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.
- **Predictions**: Activate this option to display the regression curve.
    - **Confidence intervals**: Activate this option to have confidence intervals displayed on charts (1) and (4).

Confusion plot: Activate this option to display the confusion plot which allows a synthetic visualization of the classification table. The numbers can be linked either to the width or the surface of the squares represented.

Results

XLSTAT displays a large number tables and charts to help in analyzing and interpreting the results.

Summary statistics: This table displays descriptive statistics for all the variables selected. For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

Correlation matrix: This table displays the correlations between the explanatory variables. Note that if the dependent variable is binary, the biserial correlation coefficient is used to calculate the correlation between the quantitative explanatory variables and the dependent variable.

Summary of the variables selection: Where a selection method has been chosen, XLSTAT displays the selection summary. For a stepwise selection, the statistics corresponding to the different steps are displayed. Where the number of variables varies from p to q, the best model for each number or variables is displayed with the corresponding statistics and the best model for the criterion chosen is displayed in bold.

Goodness of fit coefficients: This table displays a series of statistics for the independent model (corresponding to the case where the linear combination of explanatory variables reduces to a constant) and for the adjusted model.

- **Observations**: The total number of observations taken into account (sum of the weights of the observations);
- **Sum of weights**: The total number of observations taken into account (sum of the weights of the observations multiplied by the weights in the regression);
- **DF**: Degrees of freedom;
- **-2 Log(Like.)**: The logarithm of the likelihood function associated with the model;
- **R² (McFadden):** Coefficient, like the R²^, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model;
- **R²(Cox and Snell):** Coefficient, like the R²^, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model raised to the power 2/Sw, where Sw is the sum of weights.
- **R²(Nagelkerke):** Coefficient, like the R²^, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to ratio of the R² of Cox and Snell, divided by 1 minus the likelihood of the independent model raised to the power 2/Sw;
- **AIC**: Akaike's Information Criterion;
- **SBC**: Schwarz's Bayesian Criterion.
- **Iterations**: Number of iterations before convergence.

Test of the null hypothesis H0: Y=p0: The H0 hypothesis corresponds to the independent model which gives probability p0 whatever the values of the explanatory variables. We seek to check if the adjusted model is significantly more powerful than this model. Three tests are available: the likelihood ratio test (-2 Log(Like.)), the Score test and the Wald test. The three statistics follow a χ^2 distribution whose degrees of freedom are shown.

Type II analysis: This table is only useful if there is more than one explanatory variable. Here, the adjusted model is tested against a test model where the variable in the row of the table in question has been removed. If the probability $Pr > LR$ is less than a significance threshold which has been set (typically 0.05), then the contribution of the variable to the adjustment of the model is significant. Otherwise, it can be removed from the model.

Model parameters:

- **Binary case**: The parameter estimate, corresponding standard deviation, Wald's χ^2, the corresponding p-value and the confidence interval are displayed for the constant and each variable of the model. If the corresponding option has been activated, the "*profile likelihood* " intervals are also displayed.

- **Multinomial case**: In the multinomial case, $(J-1)*(q+1)$ parameters are obtained, where J is the number of categories and q is the number of variables in the model. Thus, for each explanatory variable and for each category of the response variable (except for the reference category), the parameter estimate, corresponding standard deviation, Wald's χ^2, the corresponding p-value and the confidence interval are displayed. The odds-ratios with corresponding confidence interval are also displayed.
- **Ordinal case**: In the ordinal case, $(J-1)+q$ parameters are obtained, where J is the number of categories and p is the number of variables in the model. Thus, for each explanatory variable and for each category of the response variable, the parameter estimate, corresponding standard deviation, Wald's χ^2, the corresponding p-value and the confidence interval are displayed.

The **equations of the model** are then displayed to make it easier to read or re-use the model.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can easily be seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

When requested, the **covariance** matrix of the parameters is then displayed.

The **marginal effects** at the point corresponding to the means of the explanatory variables are then displayed. The marginal effects are mainly of interest when compared to each other. By comparing them, one can measure the relative impact of each variable at the given point. The impact can be interpreted as the influence of a small variation of each explanatory variable, on the dependent variable. A confidence interval calculated using the Delta method is displayed. XLSTAT provides these results for both quantitative and qualitative variables, whether simple factors or interactions. For qualitative variables, the marginal effect indicates the impact of a change in category (from the first category to the category of interest).

The **predictions and residuals** table shows, for each observation, its weight, the value of the quantitative explanatory variable (if there is only one), the observed value of the dependent variable, the model's prediction, the same values divided by the weights (for the sum(binary) case), the probabilities for each category of the dependent variable, and the confidence intervals (in the binomial case).

The **influence diagnostics** table makes it possible to assess the impact of each observation on the quality of the model or on the value of the coefficients of the model. It is only displayed in the binomial and multinomial cases.

This **classification table** displays the table showing the number of well-classified and miss-classified observations for both categories. The sensitivity, specificity and the overall percentage of well-classified observations are also displayed. If a validation sample has been extracted, this table is also displayed for the validation data.

ROC curve: The ROC curve is used to evaluate the performance of the model by means of the area under the curve (AUC) and to compare several models together (see the description

section for more details).

Comparison of the categories of the qualitative variables: If one or more explanatory qualitative variables have been selected, the results of the equality tests for the parameters taken in pairs from the different qualitative variable categories are displayed.

If only one quantitative variable has been selected, the **probability analysis** table allows to see to which value of the explanatory variable corresponds a given probability of success.

Example

Tutorials on how to use logistic regression and the multinomial logit model are available on the XLSTAT Help Center:

- Logistic regression: http://www.xlstat.com/demo-log.htm

- Multinomial logit model: http://www.xlstat.com/demo-logmult.htm

- Ordinal logit model: http://www.xlstat.com/demo-logord.htm

References

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.

Log-linear regression

Use this tool to fit a log-linear regression model with three possible probability distributions (Poisson, Gamma, and Exponential).

In this section:

Description

The log-linear regression is used to model data by a log-linear combination of the model parameters and the covariates (qualitative or quantitative). Furthermore, we assume that the data (response variable) are distributed either according to a Poisson, Gamma or exponential distribution.

The log-linear regression model

Denote by Y the response variable vector, β the vector of model parameters and X the matrix of the p covariates. The first column of X is composed by a vector of 1s that deals with the intercept of the model. The log-linear model is given by:

$$E(Y|X) = e^{\beta' X}$$

According to the previous equation, we directly obtain:

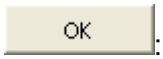
$$log[E(Y|X)] = \beta' X$$

Inference of the model parameters

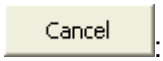
If we assume that the variables Y_i are independent from the vector of covariates X_i, the model parameters can be estimated by maximizing the likelihood. Whatever the probability distribution (Poisson, Gamma, Exponential), the likelihood function is convex and can be maximized using a Newton-Raphson algorithm.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Dependent variables:

Response variable(s): Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

Explanatory variables:

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If the variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Offset: Activate this option if you want to include an offset. This option is only available for the Poisson distribution.

Distribution: Select the probability distribution (Poisson, Gamma or Exponential).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Regression weights: Activate this option if you want to weight the influence of observations to adjust the model. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Tolerance: Enter the value of the tolerance threshold below which a variable will automatically be ignored.

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

Fixed intercept: Activate this option to set the intercept (or constant) of the model to a given value. Then enter the value in the corresponding field (0 by default).

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the Newton-Raphson algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 100.
- **Convergence**: Enter the maximum value of the evolution of the log of the likelihood from one iteration to another which, when reached, means that the algorithm is considered to

have converged. Default value: 0.000001.

Model selection: Activate this option if you want to use one of the four selection methods provided:

- **Stepwise (Forward)**: The selection process starts by adding the variable with the largest contribution to the model. If a second variable is such that its entry probability is greater than the **entry threshold value**, then it is added to the model. After the third variable is added, the impact of removing each variable present in the model after it has been added is evaluated. If the probability of the calculated statistic is greater than the **removal threshold value**, the variable is removed from the model.
- **Stepwise (Backward)**: This method is similar to the previous one but starts from a complete model.
- **Forward**: The procedure is the same as for stepwise selection except that variables are only added and never removed.
- **Backward**: The procedure starts by simultaneously adding all variables. The variables are then removed from the model following the procedure used for stepwise selection.
- **Criterion:** Choose the criterion from the following list: Likelihood, LR (likelihood ratio), Score, Wald, Akaike's AIC, Schwarz's SBC.

Validation tab:

Validation: Activate this option if you want to use a sub-sample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand,

variable labels must not be selected: the first row of the selections listed below must correspond to data.

Quantitative: Activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

Qualitative: Activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

Observations labels: Activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Variable labels: Activate this option if the first row of the data selections (explanatory variables, observations labels) includes a header.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the explanatory variables correlation matrix.

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Type III analysis: Activate this option to display the type III analysis of variance table.

Model coefficients: Activate this option to display the table of coefficients for the model.

Equation: Activate this option to display the equation for the model explicitly.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

Overdispersion test: Activate this option to test the overdispersion (only for the Poisson regression).

Charts tab:

Regression charts: Activate this option to display regression charts.

- **Confidence intervals**: Activate this option to display confidence intervals.

Prediction chart: Activate this option to display the prediction chart.

- **Confidence intervals**: Activate this option to display confidence intervals.

Results

Summary statistics: This table displays descriptive statistics for all the variables selected. For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

Correlation matrix: This table displays the correlations between the explanatory variables.

Correspondence between the categories of the response variable and the probabilities: This table shows which categories of the dependent variable have been assigned probabilities 0 and 1. It is only available for binary dependent variables.

Summary of the variable selection: Where a selection method has been chosen, XLSTAT displays the selection summary. For a stepwise selection, the statistics corresponding to the different steps are displayed. Where the best model for a number of variables varying from p to q has been selected, the best model for each number or variables is displayed with the corresponding statistics and the best model for the criterion chosen is displayed in bold.

Goodness of fit coefficients: This table displays a series of statistics for the independent model (corresponding to the case where the linear combination of explanatory variables reduces to a constant) and for the adjusted model.

- **Observations**: The total number of observations taken into account (sum of the weights of the observations).
- **Sum of weights**: The total number of observations taken into account (sum of the weights of the observations multiplied by the weights in the regression).
- **DF**: Degrees of freedom.
- **-2 Log(Like.)**: The logarithm of the likelihood function associated with the model.
- **R^2 (McFadden):** Coefficient, like the R^{2}^, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model.

- **R² (Cox and Snell):** Coefficient, like the R²^, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model raised to the power 2/Sw, where Sw is the sum of weights.
- **R² (Nagelkerke):** Coefficient, like the R²^, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to ratio of the R² of Cox and Snell, divided by 1 minus the likelihood of the independent model raised to the power 2/Sw.
- **Deviance:** Value of the deviance criterion for the adjusted model and the independent model.
- **Pearson Chi-square:** Value of the Pearson Chi-square for the adjusted model and the independent model.
- **AIC**: Akaike's Information Criterion.
- **SBC**: Schwarz's Bayesian Criterion.
- **Iterations**: Number of iterations before convergence.

Nullity test: These results allow checking whether fitted model is significantly more powerful than the independent model. Three tests are available: the likelihood ratio test (-2 Log(Like.)), the Score test and the Wald test. These three statistics follow a Chi-square distribution which degrees of freedom are shown.

Type III analysis: This table is only useful if there is more than one explanatory variable. Here, the adjusted model is tested against a test model where the variable in the row of the table in question has been removed. If the probability Pr > LR is less than a significance threshold which has been set (typically 0.05), then the contribution of the variable to the adjustment of the model is significant. Otherwise, it can be removed from the model

Model parameters: For the constant and each variable in the model, the parameter estimate, its corresponding standard deviation, the Wald's Chi- square and the corresponding p-value and the confidence interval are displayed in this table.

Example

A tutorial on how to use log-linear regression is available on XLSTAT Help Center:

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

References

Ter Berg P. (1980). On the loglinear Poisson and Gamma model. *Astin Bulletin*, **11**, 35-40.

Quantile regression

Use quantile regression to model a quantitative response variable depending on quantitative or qualitative explanatory variables. Furthermore, quantile regression makes it possible to look beyond classical regression or ANCOVA, by extending the analysis limited to expected values to the entire distribution using quantiles.

In this section:

Description

Quantile regression keeps growing in importance and interest since it was introduced by Koenker and Basset in 1978. The method popularity among the practitioners and also researchers' community is without doubt due to its peculiarity to provide them a realistic framework to perform their studies.

Indeed, by nature, quantile regression enables to work with a wide range of distributions, without being subject to any restrictions such as normality assumption, thus contrasting with usual regression.

As a consequence of that flexibility, many fields find great interest in quantile regression including Economics, Social Sciences, Environments, Biometrics and Behavioral Sciences, among others.

The main contributions on the subject may be found in the References section.

Model

As in the ANCOVA framework, the dependent variable Y is quantitative while the set of predictors X can be composed not only with quantitative variables (including interactions between quantitative variables) but also with factors (qualitative variables, interactions between qualitative variables and interactions between quantitative and qualitative variables).

Nevertheless, it's essential to keep in mind that, unlike ANCOVA, no hypotheses on the errors distribution is required.

Problem

The α -th quantile, $\alpha \in [0,1]$, is defined as the value y s.t. : $P(Y = y) = \alpha$. Introducing the cumulative distribution function F, the quantile function Q is its inverse:

$$Q(\alpha) = F^{-1}(\alpha) = \inf\{y : F(y) > \alpha\}$$

The mean μ of the random variable Y can be characterized as the value:

$$\mu = \operatorname{argmin}\{c : E[(Y - c)^2]\}$$

(1)

that minimizes the squared sum of deviations.

In the same way, the following assertion on the α -th quantile, q_α, holds :

$$q_\alpha = \operatorname{argmin}\{c : E[\rho_\alpha(Y - c)]\}$$

(2)

where ρ_α denotes the function :

$$\begin{aligned}\rho_\alpha &= [\alpha - I_{\{y<0\}}]y \\ &= [(1-\alpha)I_{\{y<0\}} + \alpha I_{\{y>0\}}]|y|\end{aligned}$$

q_α thus minimizes a weighted sum of absolute deviations.

Coming back to our context in which Y is a dependent variable and X a set of explanatory variables and considering the linear framework, the minimization problem (1) becomes:

$$\hat{\beta}_\alpha = \operatorname{argmin}\left\{\beta_\alpha : E\left[\left(Y - x_i^T \beta_\alpha\right)^2\right]\right\}$$

In the same manner, (2) turns to :

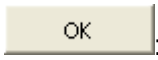
$$\hat{\beta}_\alpha = \operatorname{argmin}\left\{\beta_\alpha : E\left[\rho_\alpha\left(Y - X\beta_\alpha\right)\right]\right\}$$

where the parameters and the associated estimators depend on α .

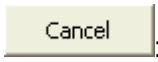
Instead of considering the conditional mean of the classical regression problem, the quantile regression problem consists in estimating conditional quantiles.

Dialog box

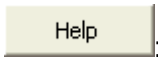
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Dependent variables:

Quantitative: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

X / Explanatory variables:

Quantitative: Select the quantitative explanatory variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Select the qualitative explanatory variables (the factors) in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

Quantile(s) :

Selection: Activate this option to work with a quantile selection. Then, select the cells containing the order α, $\alpha \in [0,1]$, of the α -th quantiles of interest in the Excel worksheet.

Process: Activate this option to get the entire quantile process. The number of calculated quantiles is specified thanks to an heuristic formula depending (in an increasing way) on the number of observations and regressors. The resulting quantile orders α are then uniformly distributed on $[0,1]$.

Options tab:

Algorithm: 3 algorithms are available to compute the quantile regression coefficients:

- **Simplex**: Select this option to compute the Barrodale and Roberts algorithm based on simplex methods.
- **Interior point**: Select this option to compute the predictor-corrector Mehrotra algorithm based on interior point methods.
- **Smoothing function**: Select this option to compute the Clark and Osborne algorithm based on an approximation of the objective function with a smooth function whose minimization provides asymptotically the same results than the initial function. This algorithm is very competitive, especially when $\frac{p}{n} > 0.05$.

Stop criterion: The selected algorithm stops as soon as one of these events occurs:

- End of the algorithm OR
- The maximum number of iterations specified in **Iterations** has been exceeded. Default value: 100. OR

- The evolution of the results from one iteration to another is inferior to the **Convergence** value, the algorithm is considered to have converged. Default value: 0.000001.

Confidence Interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Constraints: When qualitative explanatory variables have been selected, you can choose the constraints used on these variables:

- a1 = 0: Choose this option so that the parameter of the first category of each factor is set to 0.
- an = 0: Choose this option so that the parameter of the last category of each factor is set to 0.

A priori error type: Activate this option to precise, if possible, the error. Then, select the type: homogeneous error (i.i.d.), heterogeneous (i.n.i.d.) or dependent (n.i.i.d.) (for instance autocorrelated ones). This option will have an impact on the computation of the covariance matrix of the coefficients and their confidence intervals.

Validation tab:

Validation: Activate this option if you want to use a sub-sample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

Quantitative: activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

Qualitative: activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

Observations labels: activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Variable labels: Activate this option if the first row of the data selections (explanatory variables, observations labels) includes a header.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

Quantile correlations: Activate this option to display the variables quantile correlation matrix.

Covariance matrix: Activate this option to display the covariance matrix of the quantile regression coefficients.

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Model significance test: Activate this option to test the significance of the model. More precisely, the hypothesis of the complete model is tested against the hypothesis of the model made up of the intercept. 3 tests are available:

- **LR**: Activate this option to compute the Likelihood Ratio test,
- **LM**: Activate this option to compute the Lagrange Multiplier test,
- **Wald**: Activate this option to compute the Wald test.

Model equation: Activate this option to display the model equation.

Predictions and residuals: Activate this option to display the predictions and the residuals for all the observations.

Computations based on:

- **Asymptotic distribution**: Activate this option to compute the covariance matrix and the confidence intervals thanks to the theoretical asymptotic distribution of the coefficients. This computation will take into account the **A priori error type** in the **Options** tab if informed.
- **Resampling (Bootstrap)**: Activate this option to compute the empirical covariance matrix and the confidence intervals thanks to resampling (Bootstrap). Then, inform **B =** with an integer value to indicate how many samples will be simulated to compute the estimations.

· **Hall and Sheather bandwidth**: Activate this option to compute the covariance matrix and the confidence intervals using Hall and Sheather bandwidth ($O(n^{-1/3})$).

· **Bofinger bandwidth**: Activate this option to compute the covariance matrix and the confidence intervals using Bofinger bandwidth ($O(n^{-1/5})$).

Charts tab:

Regression charts: Activate this option to display the regression charts:

- **Predictions and residuals**: Activate this option to display the following charts:
- Explanatory variable versus standardized residuals: this chart is displayed only if there is one explanatory variable and if that variable is quantitative.
- Dependent variable versus standardized residuals.
- Predictions versus observed values.

Results

If the quantile process is selected in the General tab, then a global table is displayed, summing up for each computed q-th quantile, the associated coefficients value.

Charts representing the behavior of these coefficients with respect to the value of a are displayed for a better visualization of the results

If a quantile selection is chosen in the General tab, then the following results are displayed:

Summary statistics: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the dependent variables (in blue) and the quantitative explanatory variables. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

Then for each quantile, the following results are displayed:

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model:

- Observations: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.
- Sum of weights: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.
- DF: The number of degrees of freedom for the chosen model (corresponding to the error part).
- R^2_α: The determination coefficient for the model. This coefficient, whose value is between 0 and 1. Its value is defined by:

$$\begin{aligned} R^2_\alpha &= 1 - \frac{SAR_\alpha}{SAT_\alpha} \\ &= 1 - \frac{\sum_{i=1}^{n} \rho_\alpha(y_i - x_i^T \hat{\beta}_\alpha)}{\sum_{i=1}^{n} \rho_\alpha(y_i - \hat{\beta}_{0,\alpha})} \end{aligned}$$

where $\hat{y_\alpha}$ is the α-th empirical quantile of the observations of the dependent variable Y.

RAS_α is the Residual Absolute Sum of weighted differences

TAS_α is the Total Absolute Sum of weighted differences.

The R^2_α is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer R^2_a is to 1, the better is the model. The problem with the R^2_α is that it does not take into account the number of variables used to fit the model.

- Adjusted R^2_α : The adjusted determination coefficient for the model. The adjusted R^2_α can be negative if the R^2_α is near to zero. Its value is defined by:

$$\text{adjusted}{R^2}_\alpha = 1 - (1 - R^2_\alpha)\frac{W-1}{W-p-1}$$

The adjusted R^2_α is a correction to the R^2_α which takes into account the number of variables used in the model.

- $MRAS_\alpha$: The Mean Residual Absolute Sum (MRAS) is defined by :

$$MRAS_\alpha = \frac{1}{W-p} RAS_\alpha$$

- $RMRAS_\alpha$: the square Root of the Mean Residual Absolute Sum (RMRAS).

- $MAPE_\alpha$: The Mean Absolute Percentage Error is calculated as follows:

$$MAPE_\alpha = \frac{100}{W}\sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_{i,\alpha}}{y_i} \right|$$

- Cp_α : Mallows Cp coefficient is defined by :

$$Cp_\alpha = \frac{RAS_\alpha}{\hat{\sigma}_\alpha} + 2p - W$$

where RAS_α is the Residual Absolute Sum of weighted differences and $\hat{\sigma}_\alpha$ denotes the variance estimator of the residuals. The nearer Cp is to p, the less the model is biaised.

- AIC_α: Akaike's Information Criterion is defined by:

$$AIC_\alpha = W \ln\left(\frac{RAS_\alpha}{W}\right) + 2p$$

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- SBC_α: Schwarz's Bayesian Criterion is defined by:

$$SBC_\alpha = W \ln\left(\frac{RAS_\alpha}{W}\right) + \ln(W)p$$

This criterion, proposed by Schwarz (1978) is similar to the AIC and, like this, the aim is to minimize it.

- PC_α: Amemiya's Prediction Criterion is defined by :

$$PC_\alpha = \frac{(1 - R_\alpha^2)(W + p)}{W - p}$$

This criterion, proposed by Amemiya (1980) is used, like the adjusted R_α^2 to take account of the parsimony of the model.

The table of model parameters displays the estimate of the parameters, the corresponding standard error, as well as the confidence interval.

The model significance table is used to evaluate the explanatory power of the explanatory variables. The explanatory power is evaluated by comparing the fit of the final model with the fit of the rudimentary model including only a constant equal to the quantile of the dependent variable.

The predictions and residuals table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the residuals, the confidence. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the random values being larger. If the validation data have been selected, they are displayed at the end of the table.

The charts which follow show the results mentioned above. If there is only one explanatory variable in the model, the first chart displayed shows the observed values, the regression line and both types of confidence interval around the predictions. The second chart shows the standardized residuals as a function of the explanatory variable. In principle, the residuals should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a problem with the model.

Example

A tutorial on how to run a quantile regression is available on XLSTAT Help Center:

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

References

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.

Cubic splines

This tool allows to fit a cubic spline to using a set of nodes defined by the user.

In this section:

Description

A cubic spline is defined as a piecewise function of polynomials of degree 3. Cubic splines are used in interpolation problems where they are preferred to usual polynomial interpolation methods, because they allow a compromise between the smoothness of the curve and the degree of the polynomials.

Cubic splines

A cubic spline S is a piecewise function defined on an interval $[a, b]$ divided into K intervals $[x_{i-1}, x_i]$ such that

$$a = x_0 < x_1 < \ldots < x_{K-1} < x_K = b$$

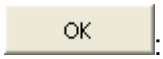
and it is defined by P_i the polynomial of degree 3 on the interval $[x_{i-1}, x_i]$. The spline S is given by:

$$\begin{cases} S(t) = P_1(t) & \text{if} \quad t \in [x_0, x_1] \\ \vdots \\ S(t) = P_K(t) & \text{if} \quad t \in [x_{K-1}, x_K] \end{cases}$$

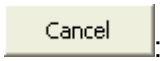
The calculation of the coefficients of the cubic spline involves the derivatives of the polynomials (for further details see Guillod, 2008).

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y: Select the data corresponding to the ordinates. If the name of the variable is available in the first position of the data, make sure you activate the "Variable labels" option.

X: Select the data that correspond to the abscissa. If the name of the variable is available in the first position of the data, make sure you activate the "Variable labels" option.

Groups: Check this option to select the values which correspond to the identifier of the group to which each observation belongs. On the chart, the color of the point depends on the group.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Optionstab:

Data as nodes: Activate this option to use the data as nodes for the cubic spline.

Number of nodes: Activate this option to select the number of nodes. This nodes are then equally distributed.

Select the nodes coordinates: If the option is enabled, you have to select the range containing the coordinates of the nodes.

Prediction tab:

Prediction: Activate this option if you want to select data to use them in prediction mode. If you activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

Observations: Select the variables for prediction. The first row must not include variable labels.

Observations labels: Activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the correlation matrix.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

Charts tab:

Spline curve: Activate this option to display the spline curve

- **Predictions and residuals**: Activate this option to display the following charts.

(1) Explanatory variable versus standardized residuals: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

(2) Dependent variable versus standardized residuals.

(3) Predictions for the dependent variable versus the dependent variable.

(4) Bar chart of standardized residuals.

- **Splines on the same chart**: This option is only available when groups have been selected. Activate this option if you want to display all splines on a single chart.

Results

Summary statistics: This table displays the descriptive statistics for each element

Coefficients of the cubic spline: for each interval the coefficients of the cubic spline are given in a table.

Example

An example of the use of cubic splines is available on XLSTAT Help Center:

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

References

Guillod T. (2008). Interpolations, courbes de Bézier et B-splines. *Bulletin de la société des Enseignants Neuchatelois de Sciences*, **34**.

Nonparametric regression

This tool carries out two types of nonparametric regression: *Kernel regression* and LOWESS regression.

In this section:

Description

Parametric regression can be used when the hypotheses about the more classical regression methods cannot be verified or when we are mainly interested in only the predictive quality of the model and not its structure.

Kernel regression

Kernel regression is a modeling tool which belongs to the family of smoothing methods. Unlike linear regression which is both used to explain phenomena and for prediction (understanding a phenomenon to be able to predict it afterwards), *Kernel regression* is mostly used for prediction. The structure of the model is variable and complex, the latter working like a filter or black box. There are many variations of *Kernel regression* in existence.

As with any modeling method, a learning sample of size n_{learn} is used to estimate the parameters of the model. A sample of size n_{valid} can then be used to evaluate the quality of the model. Lastly, the model can be applied to a prediction sample of size n_pred, for which the values of the dependent variable Y are unknown.

The first characteristic of *Kernel Regression* is the use of a **kernel** function, to weigh the observations of the learning sample, depending on their "distance" from the predicted observation. The closer the values of the explanatory variables for a given observation of the learning sample are to the values observed for the observation being predicted, the higher the weight. Many kernel functions have been suggested. XLSTAT includes the following kernel functions: Uniform, Triangle, Epanechnikov, Quartic, Triweight, Tricube, Gaussian, and Cosine.

The second characteristic of *Kernel regression* is the **bandwidth** associated to each variable. It is involved in calculating the kernel and the weights of the observations, and differentiates or rescales the relative weights of the variables while at the same time reducing or augmenting the impact of observations of the learning sample, depending on how far they are from the

observation to predict. The term *bandwidth* refers to the filtering methods. The lower the bandwidth, the fewer will be the number of observations to influence the prediction.

Example: let Y be the dependent variable, and $(X_1, X_2, \ldots, X_k)$ the k explanatory variables. For the prediction of y_i from observation i $(1 \leq i \leq n_{valid})$, given the observation j $(1 \leq j \leq n_{learn})$, the weight determined using a Gaussian kernel, with a bandwidth fixed to h_l for each of the X_l variables $(l = 1 \ldots k)$, is given by:

$$w_{ij} = \frac{1}{\left(\sqrt{2\pi}\right)^k \prod_{l=1}^{k} h_l} \exp\left(-\sum_{l=1}^{k}\left(\frac{x_{jl} - x_{il}}{h_l}\right)^2\right)$$

The third characteristic is the polynomial degree used when fitting the model to the observations of the learning sample. In the case where the polynomial degree is 0 (constant polynomial), the Nadaraya-Watson formula is used to compute the ith prediction:

$$y_i = \frac{\sum_{j=1}^{n_{learn}} w_{ij} y_j}{\sum_{j=1}^{n_{learn}} w_{ij}}$$

For the constant polynomial, the explanatory variables are only taken into account for computing of the weight of the observations in the learning sample. For higher polynomial degrees (experience shows that higher orders are not necessary and XLSTAT works with polynomials of degrees 0 to 2), the variables are used in calculating a polynomial model. Once the model has been fitted, it is applied to the validation or prediction sample in order to estimate the values of the dependent variable.

Once the parameters of the model have been estimated, the prediction value is calculated using the following formulae:

1. Degree 1: $y_i = a_0 + \sum_{l=1}^{k} a_l x_{il}^l$
2. Degree 2: $y_i = a_0 + \sum_{l=1}^{k} a_l x_{il}^l + \sum_{l=1}^{k}\sum_{m=1}^{k} b_{lm} x_{il} x_{im}$

Notes:

- Before we estimate the parameters of the polynomial model, the observations of the learning sample are previously weighted using the Nadaraya-Watson formula.
- For a 1st or 2nd order model, for each observation of the validation and prediction samples, the polynomial parameters are estimated. This makes *Kernel Regression* a numerically intensive method.

Two strategies are suggested in order to restrict the size of the learning sample taken into account for the estimation of the parameters of the polynomial:

- Moving window: to estimate y_i, we take into account a fixed number of observations previously observed. Consequently, with this strategy, the learning sample evolves at each step.
- k nearest neighbors: this method, complementary to the previous, restricts the size of the learning sample to a given value k.

Details of the kernel functions:

The weight w_{ij} computed for observation j, for the estimation of prediction y_i, is defined as:

$$W_{ij} = \prod_{l=1}^{k} \frac{K(u_{ijl})}{h_l} \text{ where } u_{ijl} = \frac{x_{il} - x_{jl}}{h_l}$$

and K is a kernel function. The kernel functions available in XLSTAT are:

- Uniform: the kernel function is defined as:

$$K(u) = \frac{1}{2}\mathbb{I}_{|u|\leq 1}$$

- Triangle: the kernel function is defined as:

$$K(u) = (1 - |u|)\,\mathbb{I}_{|u|\leq 1}$$

- Epanechnikov: the kernel function is defined as:

$$K(u) = \frac{3}{4}\left(1 - u^2\right)\mathbb{I}_{|u|\leq 1}$$

- Quartic: the kernel function is defined as:

$$K(u) = \frac{15}{16}(1 - u^2)^2\mathbb{I}_{|u|\leq 1}$$

- Triweight: the kernel function is defined as:

$$K(u) = \frac{35}{32}(1 - u^2)^3\mathbb{I}_{|u|\leq 1}$$

- Tricube: the kernel function is defined as:

$$K(u) = \left(1 - |u|^3\right)^3\mathbb{I}_{|u|\leq 1}$$

- Gaussian: the kernel function is defined as:

$$K(u) = \frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$$

- Cosine: the kernel function is defined as:

$$K(u) = \frac{\pi}{4} cos\left(\frac{\pi}{2}u\right) \mathbb{I}_{|u|\leq 1}$$

LOWESS regression

LOWESS regression (Locally weighted regression and smoothing scatter plots) was introduced by Cleveland (1979) in order to create smooth curves through scattergrams. New versions have since been perfected to increase the robustness of the models. LOWESS regression is very similar to *Kernel regression* as it is also based on polynomial regression and requires a kernel function to weight the observations.

The LOWESS algorithm can be described as follows: for each point i to predict:

1 - First, the Euclidean distances $d(i,j)$ between the observations i and j are computed. The fraction f of the N closest observations to observation i are selected. The weight of the selected points are selected using the Tricube kernel and the following distance:

$$D(i,j)\frac{d(i,j)}{Max_j\left(d(i,j)\right)}$$

$$Weight(j) = Tricube\left(D(i,j)\right)$$

2 - A regression model is then fitted, and a prediction is computed for observation i.

For the robust LOWESS regression, additional computations are performed:

3 - The weights are computed again using the following distance:

$$D^{'}(i,j) = \frac{|r(j)|}{6Median_j\left(|r(j)|\right)}$$

where $r(j)$ is the residual corresponding to observation j after the previous step.

Using the Quartic kernel:

$$Weight(j) = Quartic\left(D^{'}(i,j)\right)$$

4 - The regression is then fitted again using the new weights.

5 - Steps 3 and 4 are performed a second time. A final prediction is then computed for observation i.

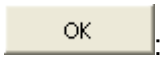
Notes:

- The only input parameters apart from the observations for the method are the f fraction of nearest individuals (in % in XLSTAT) and the polynomial degree.

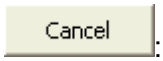
- *Robust LOWESS regression* is about three times more time consuming than LOWESS regression.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Dependent variables:

Quantitative: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

X / Explanatory variables:

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Method: Choose the type of nonparametric regression to use (see description).

Polynomial degree: enter the order of the polynomial if the LOWESS regression or a polynomial method is chosen.

Options tab:

Learning samples:

- **Moving window**: choose this option if you want the size of the learning sample to be constant. You need to enter the size S of the window. In that case, to estimate $Y(i+1)$, the observations $i-S-1$ to i will be used, and the first observation XLSTAT will be able to compute a prediction for, is the $S+1$'th observation.
- **Expanding window**: choose this option if you want the size of the learning sample to be expanding step by step. You need to enter the initial size S of the window. In that case, to estimate $Y(i+1)$, observations 1 to i will be used, and the first observation XLSTAT will be able to compute a prediction for, is the $S+1$'th observation.
- **All**: the learning and validation samples are identical. This method has no interest for prediction, but it is a way to evaluate the method in case of perfect information.

K nearest neighbors:

- **Rows**: the k points retained for the analysis are k points which are the closest to the point to predict, for a given bandwidth and a given kernel function. k is the value to enter here.
- **%**: the points retained for the analysis are the closest to the point to predict and represent x% of the total learning sample available, where x is the value to enter.

Tolerance: Enter the value of the tolerance threshold below which a variable will automatically be ignored.

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Kernel: the kernel function that will be used. The possible options are: Uniform, Triangle, Epanechnikov, Quartic, Triweight, Tricube, Gaussian, Cosine. A description of these functions is available in the description section.

Bandwidth: XLSTAT allows you to choose a method for automatically computing the bandwidths (one per variable), or you can fix them. The different options are:

- **Constant**: the bandwidth is constant and equal to the fixed value. Enter the value of the bandwidth.
- **Fixed**: the bandwidth is defined in a vertical range of cells in an Excel sheet, which you need to select. The cells must be equal to the number of explanatory variables, and in the same order as the variables.
- **Range**: the value h_l of the bandwidth for each variable X_l is determined by the following formula:

$$h_l = Max\left(x_{il}\right)_{i=1,\ldots,n_{learn}} - Min\left(x_{il}\right)_{i=1,\ldots,n_{learn}}$$

- **Standard deviation**: the value h_l of the bandwidth for each explanatory variable is equal to the standard deviation of the variable computed on the learning sample.

Validation tab:

Validation: Activate this option if you want to use a sub-sample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

Quantitative: Activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

Qualitative: Activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

Observations labels: activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Variable labels: Activate this option if the first row of the data selections (explanatory variables, observations labels) includes a header.

Missing data tab:

These options are available only for PCR and OLS regression. With PLS regression, the missing data are automatically handled by the algorithm.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the explanatory variables correlation matrix.

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

Charts tab:

Histograms: Activate this option to display a histogram behind the kernel density curve. Use the following options to control the display of the histogram:

- **Intervals**: Choose one of the following options to define the intervals for the histogram:
- **Number**: Choose this option to enter the number of intervals to create.
- **Width**: Choose this option to define a fixed width for the intervals.
- **User defined**: Select a column containing in increasing order the lower bound of the first interval, and the upper bound of all the intervals.
- **Minimum**: Activate this option to enter the value of the lower value of the first interval. This value must be lower or equal to the minimum of the series.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the selected variables. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables. For qualitative variables, the categories with their respective frequencies and percentages are displayed.

Correlation matrix: This table displays the correlations between the selected variables.

Goodness of fit coefficients: This table shows the following statistics:

- The determination coefficient R^2;
- The sum of squares of the errors (or residuals) of the model (SSE or SSR respectively);
- The means of the squares of the errors (or residuals) of the model (MSE or MSR);
- The root mean squares of the errors (or residuals) of the model (RMSE or RMSR).

Predictions and residuals: Table giving for each observation the input data, the value predicted by the model and the residuals.

Charts:

If only one quantitative explanatory variable or temporal variable has been selected ("As a function of time" option in the "Charts" tab in the dialog box), the first chart shows the data and the curve for the predictions made by the model. If the "As a function of X1" option has been selected, the first chart shows the observed data and predictions as a function of the first explanatory variable selected. The second chart displayed is the bar chart of the residuals.

Example

A tutorial on Kernel regression is available on the XLSTAT Help Center:

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

References

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.

Nonlinear regression

Use this tool to fit data to any linear or non-linear function. The method used is least squares. Either pre-programmed functions or functions added by the user may be used.

In this section:

Description

Nonlinear regression is used to model complex phenomena which cannot be handled by linear models. XLSTAT provides preprogrammed functions from which the user may be able to select the model which describes the phenomenon to be modeled.

When the model required is not available, the user can define a new model and add it to their personal library. To improve the speed and reliability of the calculations, it is recommended to add derivatives of the function for each of the parameters of the model.

In order to calculate these parameters, the Levenberg-Marquardt algorithm is used.

XLSTAT allows you to model several functions at the same time. You can then choose to display only the results of the best model (based on AIC), or those of all models.

Levenberg-Marquardt algorithm

The Levenberg-Marquardt algorithm is an iterative technique that locates the minimum of a multivariate function. This technique is recommended for non-linear least squares regression problems.

Consider the nonlinear model $Y = f(X, \Theta)$, where θ is the parameter vector of size px_1, X is the vector of the explanatory variables, and f is a function of X and Θ. The goal is to find the least squares Θ' estimate of Θ such that Θ' minimizes the function f. We therefore try to minimize the following function:

$$g(\Theta) = \sum_{i=1}^{m} \left(Y_i - f(\Theta, X_i)\right)^2) \cdot$$

The procedure is iterative, starting from an initial parameter, if possible close to the final solution, and applying the following routine:

$$\Theta_{j+1} = \Theta_j - (J'J + \lambda D)^{-1} J'(Y - f(\Theta, X_i)),$$

where J is the Jacobian matrix, and D is a diagonal matrix to adjust the λ damping parameter.

When derivatives are not specified, a finite difference approximation is used to estimate them.

Global fitting and shared parameters:

In XLSTAT, you have the possibility to adjust multiple variables at the same time. There are two ways to do this:

- The first one is to have one column for each dependent variable Y.
- The second one is to have one column containing all the Y variables to be adjusted and another column which contains the group indices making it possible to identify each Y. With this option, there you can choose a set of shared parameters that apply to a set of curves.

Adding a function to the library of user-defined functions

Syntax:

The parameters of the function must be written as pr_1, pr_2, etc..

The explanatory variables must be represented as X_1, X_2, etc..

Excel functions can be used: Exp(), Sin(), Pi(), Max(), etc.

Example of a function: pr_1 * Exp(pr_2 + pr_3 * X_1 + pr_4 * X_2)

File containing function definitions:

The library of user functions is held in the file Models.txt in the user directory defined during installation or by using the options XLSTAT dialog box. The library is built as follows:

Row 1: number of functions defined by user.

Row 2: N1= number of parameters in function 1.

Row 3: function 1 definition.

Rows 4 to (3 + N1): derivatives definition for function 1.

Row 4+N1: N2= number of parameters in function 2.

Row 5+N1: function 2 definition…

When the derivatives have not been supplied by the user, "Unknown" replaces the derivatives of the function.

You can modify manually the items of this file but you should be cautious not to make an error.

R² for nonlinear regression

In their work, Spiess and Neumeyer performed various simulations demonstrating that using R^2 to evaluate the fit of nonlinear models can lead to incorrect conclusions.

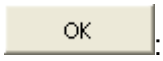
In particular, they observed that the R^2 tends to be uniformly high, whether the models are very poor or excellent. Additionally, R^2 and adjusted R^2 do not necessarily improve with higher quality nonlinear models.

Therefore, we now advise against relying on this statistic and have stopped displaying it in XLSTAT.

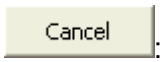
To compare models, you can look at the RMSE statistic, the lower it is, the better the model fits the data. Spiess and Neumeyer also showed that AIC and corrected AIC perform better than R^2, so it is also possible to look at these statistics to validate the model.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Y / Dependent variables:

Quantitative: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

X / Explanatory variables:

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

Groups: Activate this option if you want to include a group variable. This allows you to adjust simultaneously several variables. Then select the corresponding variable on the Excel sheet. If the label of the variables has been selected, please check that the "Variable labels" option is activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. If the "Variable labels" option is activated you need to include a header in the selection.

Functions tab:

Built-in function: Activate this option to fit one of the functions available from the list of built-in functions to the data. Select a function from the list.

Edit: Click this button to display the active built-in function in the "Function: Y=" field. You can then copy the function to afterwards change it to create a new function or the derivatives of a new function.

User defined functions: Activate this option to fit one of the functions available from the list of user-defined functions to the data, or to add a new function.

Choose a model: Activate this option to fit a function to the data and display its results.

Choose a model among several: Activate this option to fit several functions to the data and display the results of the best model, based on the AIC.

Choose several models: Activate this option to fit several functions to the data and display their results.

Delete: Click this button to delete the active function from the list of user-defined functions.

Add: Click this button to add a function to the list of user-defined functions. You must then enter the function in the "**Function: Y=**" field, then, if you want and given that it will speed up the calculations and enable the standard deviations of the parameters to be obtained, you can select the derivatives of the function for each of the parameters. To do this, activate the "**Derivatives**" option, then select the derivatives in an Excel worksheet.

Derivatives: These will speed up the calculations and enable the standard deviations of the parameters to be obtained.

Details: Click this button to get more information about the selected built-in function.

Inhibitor concentration: If you select at least one of the following 4 functions: "Competitive inhibition", "Non-competitive inhibition", "Uncompetitive inhibition", or "Mixed model inhibition", you must select the inhibitor concentration. A different concentration is required per group.

Note: the description section contains information on defining user functions.

Options tab:

Initial values: Activate this option to give XLSTAT a starting point. Select the cells which correspond to the initial values of the parameters. The number of rows selected must be the same as the number of parameters.

Parameters bounds: Activate this option to give XLSTAT a possible region for all the parameters of the model selected. You must them select a two- column range, the one on the left being the lower bounds and the one on the right the upper bounds. The number of rows selected must be the same as the number of parameters.

Parameters labels: Activate this option if you want to specify the names of the parameters. XLSTAT will display the results using the selected labels instead of using generic labels pr1, pr2, etc. The number of rows selected must be the same as the number of parameters.

Shared parameters: Activate this option if you want to add shared parameters to the model. This option is only available if you have added a group variable. The shared parameters will then have the same parameter value for all the fitted groups.

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 200.
- **Convergence**: Enter the maximum value of the evolution in the Sum of Squares of Errors (SSE) from one iteration to another which, when reached, means that the algorithm is

considered to have converged. Default value: 0.00001.

Validation tab:

Validation: Activate this option if you want to use a sub-sample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

Quantitative: Activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

Qualitative: Activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

Observations labels: activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Variable labels: Activate this option if the first row of the data selections (explanatory variables, observations labels) includes a header.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the explanatory variables correlation matrix.

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Model parameters: Activate this option to display the values of the parameters for the model after fitting.

Equation of the model: Activate this option to display the equation of the model once fitted.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

Display charts:

- **Data and predictions**: Activate this option to display the chart of observations and the curve for the fitted function.
- **Residuals**: Activate this option to display the residuals as a bar chart.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected: the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

Correlation matrix: This table displays the correlations between the selected variables.

Summary table: This table is displayed if you have chosen to adjust several models. In this table are displayed the goodness of fit coefficients for each of the adjusted models. Following this table is displayed a graph of the AIC of each model.

Goodness of fit coefficients: This table shows the following statistics:

- The number of observations;
- The degrees of freedom (DF);
- The sum of squares of the errors (or residuals) of the model (SSE or SSR respectively);
- The means of the squares of the errors (or residuals) of the model (MSE or MSR);
- The root mean squares of the errors (or residuals) of the model (RMSE or RMSR);
- **AIC**: Akaike's Information Criterion;

- **AICC**: Akaike's Information Criterion Corrected;
- **Iterations**: Number of iterations before convergence.

Model parameters: This table gives the value of each parameter after fitting to the model, its associated standard deviation as well as the 95% confidence interval.

Predictions and residuals: This table gives for each observation the input data, the value predicted by the model and the residuals. It is followed by the **equation** of the model.

Charts:

If only one quantitative explanatory variable has been selected, the first chart represents the data and the curve for the chosen function. In this chart, you can display the 95% confidence interval curves as well as the 95% prediction interval curves. The confidence interval allows you to assess the fitted value for the observed values of the variables, while the prediction interval gives a range of values around which a future observation of the dependent variable can be expected.

The second chart displayed is the bar chart of the residuals.

Example

Tutorials showing how to run a nonlinear regression are available on the XLSTAT Help Center on the following pages:

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

http://www.xlstat.com/demo-nonlin2.htm

References

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.

Spiess, Andrej-Nikolai, Natalie Neumeyer (2010). An evaluation of R^2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. BMC Pharmacology.

Two-stage least squares regression

Use this tool to analyze your data with a two-stage least squares regression.

In this section:

Description

The two-stage least squares method is used to handle model with endogenous explanatory variables in a linear regression framework. An endogenous variable is a variable which is correlated with the error term in the regression model. Using endogenous variable is in contradiction with the linear regression assumptions. This kind of variable can be encountered when variable are measured with error.

The general principle of the two-stage least squares approach is to use instrumental variables uncorrelated with the error term to estimate the model parameters. These instrumental variables are correlated to the endogenous variables but not with the error term of the model.

Denote by y the quantitative dependent variable, X_1 the matrix of $p1$ endogenous explanatory variables, X_2 the matrix of $p2$ exogenous explanatory variables (not correlated to the error term) ($p = p1 + p2$) and Z the matrix of q instrumental variables. The structural equations of the model are given by:

$$\begin{cases} y = X_1\beta_1 + X_2\beta_2 + \epsilon \\ X_1 = Z\gamma + \delta \end{cases}$$

where β_1 and β_2 are the parameters respectively associated to X_1 and X_2. The variables ϵ and δ are the disturbances with zero means.

According to the estimation technique developed by Theil (1953a, 1953b), the estimate of the parameter $\beta = (\beta_1, \beta_2)$ is given by:

$$\hat{\beta} = (X'\Omega X)^{-1} X'\Omega X y$$

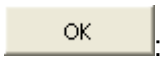
where $\Omega = Z(Z'Z)^{-1}Z'$ is the projection matrix.

XLSTAT enables you to take into account endogenous, exogenous and instrumental variables. Endogenous and exogenous variables should be selected as explanatory variables and

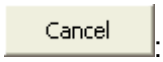
instrumental and exogenous variables should be selected as instrumental variables (exogenous variables should be selected in both selections).

Dialog box

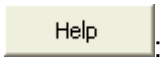
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Dependent variables:

Quantitative: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

X / Explanatory variables:

Quantitative: Select the quantitative explanatory variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated. Exogenous and endogenous variables should be selected here.

Z / Instrumental variables:

Quantitative: Select the quantitative instrumental variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the

"Variable labels" option has been activated. All the exogenous variables must be selected as instrumental variables.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

Regression weights: Activate this option if you want to carry out a weighted least squares regression. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Null intercept: Activate this option to set the constant of the model to 0.

Tolerance: Activate this option to prevent the OLS regression calculation algorithm taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

Validation tab:

Validation: Activate this option if you want to use a sub-sample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

X / Explanatory variables: Select the quantitative explanatory variables. The first row must not include variable labels. Only exogenous and endogenous variables should be selected here.

Observations labels: activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

- **Check for each Y separately**: Choose this option to remove the observations with missing data in the select Y (dependent) variables, only if the Y of interest has a missing data.
- **Across all Ys**: Choose this option to remove the observations with missing data in the Y (dependent) variables, even if the Y of interest has no missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the correlation matrix for quantitative variables (dependent or explanatory).

Analysis of variance: Activate this option to display the analysis of variance table.

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

Charts tab:

Regression charts: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.
- **Predictions and residuals**: Activate this option to display the following charts.

(1) Line of regression: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

(2) Explanatory variable versus standardized residuals: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

(3) Dependent variable versus standardized residuals.

(4) Predictions for the dependent variable versus the dependent variable.

(5) Bar chart of standardized residuals.

Confidence intervals: Activate this option to have confidence intervals displayed on charts (1) and (4).

Results

Summary statistics: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the dependent variables (in blue) and the quantitative explanatory variables. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

Correlation matrix: This table is displayed to give you a view of the correlations between the various variables selected.

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.
- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).
- **R²**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} w_i \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2}, \quad \text{where } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} w_i y_i ,$$

The R² is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer R² is to 1, the better is the model. The problem with the R² is that it does not take into account the number of variables used to fit the model.

- **Adjusted R²**: The adjusted determination coefficient for the model. The adjusted R² can be negative if the R² is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}$$

The adjusted R² is a correction to the R² which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W - p^*}\sum_{i=1}^{x} w_i (y_i - \hat{y}_i)^2$$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.
- **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows:

$$MAPE = \frac{100}{W}\sum_{i=1}^{n} w_i \left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

- **DW**: The Durbin-Watson statistic is defined by:

$$DW = \frac{\sum_{i=2}^{x}[(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{x} w_i(y_i - \hat{y}_i)^2}$$

This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- **Cp**: Mallows Cp coefficient is defined by:

$$Cp = \frac{SSE}{\hat{\sigma}} + 2p^* - W$$

where SSE is the sum of the squares of the errors for the model with p explanatory variables and $\hat{\sigma}$ is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to p*, the less the model is biased.

- **AIC**: Akaike's Information Criterion is defined by:

$$AIC = W \ln(\frac{SSE}{W}) + 2p^*$$

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz's Bayesian Criterion is defined by:

$$SBC = W \ln(\frac{SSE}{W}) + \ln(W)p^*$$

This criterion, proposed by Schwarz (1978) is similar to the AIC, and the aim is to minimize it.

- **PC**: Amemiya's Prediction Criterion is defined by:

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}$$

This criterion, proposed by Amemiya (1980) is used, like the adjusted R² to take account of the parsimony of the model.

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

The **parameters of the model** table displays the estimate of the parameters, the corresponding standard error, the Student's t, the corresponding probability, as well as the confidence interval.

The **equation of the model** is then displayed to make it easier to read or re-use the model.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of normalized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the residuals and the confidence intervals with the fitted prediction. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the random values being larger. If the validation data have been selected, they are displayed at the end of the table.

The **charts** which follow show the results mentioned above. If there is only one explanatory variable in the model, the first chart displayed shows the observed values, the regression line and both types of confidence interval around the predictions. The second chart shows the normalized residuals as a function of the explanatory variable. In principle, the residuals should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a problem with the model.

The **three charts** displayed next show respectively the evolution of the standardized residuals as a function of the dependent variable, the distance between the predictions and the observations (for an ideal model, the points would all be on the bisector), and the standardized residuals on a bar chart. The last chart quickly shows if an abnormal number of values are outside the interval]-2, 2[given that the latter, assuming that the sample is normally distributed, should contain about 95% of the data.

If you have selected the data to be used for calculating **predictions on new observations**, the corresponding table is displayed next.

Example

A tutorial on the two-stage least square approach is available on the XLSTAT Help Center:

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

References

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.

PLS/PCR Regression

Use this module to model and predict the values of one or more dependent quantitative or qualitative variables using a linear combination of one or more explanatory quantitative and/or qualitative variables, without facing the constraints of OLS (ordinary least square regression) on the number of variables versus the number of observations.

In this section:

Description

The regression methods available in this module have the common characteristic of generating models that involve linear combines of explanatory variables. The difference between the three method lies on the way the correlation structures between the variables are handled.

PCR Regression

PCR (Principal Components Regression) can be divided into three steps: we first run a PCA (Principal Components Analysis) on the table of the explanatory variables, then we run an OLS regression on the selected components, then we compute the parameters of the model that correspond to the input variables.

PCA allows to transform an X table with n observations described by variables into an S table with n scores described by q components, where q is lower or equal to p and such that $(S'S)$ is invertible. An additional selection can be applied on the components so that only the r components that are the most correlated with the Y variable are kept for the OLS regression step. We then obtain the R table.

The OLS regression is performed on the Y and R tables. In order to circumvent the interpretation problem with the parameters obtained from the regression, XLSTAT transforms the results back into the initial space to obtain the parameters and the confidence intervals that correspond to the input variables.

PLS Regression

This method is quick, efficient and optimal for a criterion based on covariances. It is recommended in cases where the number of variables is high, and where it is likely that the explanatory variables are correlated.

The idea of PLS regression is to create, starting from a table with n observations described by p variables, a set of h components with $h < p$. The method used to build the components differs from PCA, and presents the advantage of handling missing data. The determination of the number of components to keep is usually based on a criterion that involves a cross-validation. The user may also set the number of components to use.

Some programs differentiate PLS1 from PLS2. PLS1 corresponds to the case where there is only one dependent variable. PLS2 corresponds to the case where there are several dependent variables. The algorithms used by XLSTAT are such that the PLS1 is only a particular case of PLS2.

In the case of the OLS and PCR methods, if models need to be computed for several dependent variables, the computation of the models is simply a loop on the columns of the dependent variables table Y. In the case of PLS regression, the covariance structure of Y also influences the computations.

The equation of the PLS regression model with h components writes:

$$\begin{aligned} Y &= T_h C_h' + E_h \\ &= X W_k^* C_h' + E_h \\ &= X W_h (P_h' W_h)^{-1} C_h' + E_h \end{aligned}$$

where Y is the matrix of the dependent variables, X is the matrix of the explanatory variables. T_h, C_h, W_h^*, W_h and P_h, are the matrices generated by the PLS algorithm, and E_h is the matrix of the residuals.

The matrix B of the regression coefficients of Y on X, with h components generated by the PLS regression algorithm is given by:

$$B = W_h (P_h' W_h)^{-1} C_h'$$

Note: The PLS regression leads to a linear model as the OLS and PCR do.

Notes:

The three methods give the same results if the number of components obtained from the PCA (in PCR) or from the PLS regression is equal to the number of explanatory variables.

The components obtained from the PLS regression are built so that they explain as well as possible Y, while the components of the PCR are built to describe X as well as possible. XLSTAT allows partly compensating this drawback of the PCR by allowing the selection of the components that are the most correlated with Y.

PLS Discriminant Analysis

PLS regression can be adapted to fit discriminant analysis. The PLS discriminant analysis uses the PLS algorithm to explain and predict the membership of observations to several classes using quantitative or qualitative explanatory variables. XLSTAT-PLS uses the PLS2 algorithm applied on the full disjunctive table obtained from the qualitative dependent variable.

PLS discriminant analysis can be applied in many cases when classical discriminant analysis cannot be applied. For example, when the number of observations is low and when the number of explanatory variables is high. When there are missing values, PLS discriminant analysis can be applied on the data that is available. Finally, as PLS regression, it is adapted when multicollinearity between explanatory variables is high.

As many models as categories of the dependent variable are obtained. An observation is associated to the category that has an equation with the highest value.

Let K be the number of categories of the dependent variable Y. For each category a_1, ..., a_K, an equation of the model is obtained:

$$F(y_i, a_k) = b_0 + \sum_{j=1}^{p} b_i x_{ij}$$

With a_k being a category of the dependent qualitative variable, b_0 being the intercept of the model associated to a_k, p being the number of explanatory variables and b_i being the coefficients of the same model.

Observation i is associated to class k if:

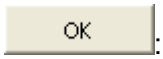
$$k^* = \arg\max_k F(y_i, a_k)$$

PLS discriminant analysis offers an interesting alternative to classical linear discriminant analysis.

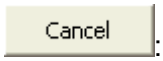
The output mixes the outputs of the PLS regression with classical discriminant analysis outputs such as confusion matrix.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 ![Select_list.png][image-9]{: width="26" height="25"} ![Select_file.png][image-10]{: width="26" height="25"} : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files ![Select_file_choosefile.png][image-11]{: width="26" height="25"}.

General tab:

Y / Dependent variables:

Quantitative/Qualitative: Select the dependent variable(s). The data must be numerical except in the case of the PLS-DA where they can be nominal and are anyway considered as categorical. If the "Variable labels" option is activated make sure that the headers of the variables have also been selected.

X / Explanatory variables:

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables. Then select the corresponding data. The data must be numerical. If the "Variable labels" option is activated make sure that the headers of the variables have also been selected.

Qualitative: Activate this option if you want to include one or more qualitative explanatory variables. Then select the corresponding data. Whatever their Excel format, the data are considered as categorical. If the "Variable labels" option is activated make sure that the headers of the variables have also been selected.

Method: Choose the regression method you want to use:

- **PLS-R**: Activate this option to compute a Partial Least Squares regression.
- **PLS-DA**: Activate this option to compute a Partial Least Squares Discriminant Analysis.
- **PCR**: Activate this option to compute Principal Components Regression.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Observation weights: Activate this option if you want to weight the observations. If you do not activate this option, the weights are considered to be equal to 1. The weights must be greater or equal to 0 and they must be integer values. Setting a case weight to 2 is equivalent to repeating twice the same observation. If the "Variable labels" option is activated, make sure that the header of the selection has also been selected.

Regression weights: This option is active only with PCR and OLS regression. Activate this option if you want to run a weighted least squares regression. If you do not activate this option, the regression weights are considered to be equal to 1. The weights must be greater or equal to 0. If the "Variable labels" option is activated make sure that the header of the selection has also been selected.

Options tab:

Common options:

Confidence interval (%): Enter the size in % of the confidence interval that is used for the various tests, parameters and predictions. Default value: 95.

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Options for PLS regression:

Stop conditions:

- **Automatic**: Activate this option so that XLSTAT automatically determines the number of components to keep.
- **Fixed number**: Activate this option to set the number of components to take into account in the model.
- **Qi² threshold**: Activate this option to set the threshold value of the Qi² criterion used to determine if the contribution of a component is to any dependent variable significant or not. The default value is 0.0975 which corresponds to 1-0.95².
- **Qi² threshold (global)**: Activate this option to set the threshold value of the Qi² criterion used to determine if the contribution of a component to all dependent variables is significant or not. The default value is 0.0975 which corresponds to 1-0.95².

- **Qi² improvement**: Activate this option to set the threshold value of the Qi² improvement criterion used to determine if the contribution of a component is significant or not. The default value is 0.05 which corresponds to a 5% improvement. This value is computed as follows:

$$Q^2(h)\,\mathrm{Imp} = \frac{Q^2(h) - Q^2(h-1)}{Q^2(h-1)}$$

- **Minimum Press**: Activate this option so that the number of components used in the model corresponds to the model with the minimum Press (predicted residual error sum of squares) statistic.

X / Explanatory variables:

- **Center**: Activate this option is you want to center the explanatory variables before starting the calculations.

- **Reduce**: Activate this option is you want to standardize the explanatory variables before starting the calculations.

Algorithm: The difference between the two approaches can only be seen if a jackknife option was selected for cross validation, and if the data are centered or standardized.

- **Fast**: Activate this option to use a faster algorithm. The algorithm avoids recentering or restandardizing the jackknife training sets.

- **Precise**: Activate this option to use the slower but more precise algorithm. The algorithm recenters or restandardizes the jackknife training sets.

Cross validation:

- **None**: Activate this option so that XLSTAT does not run use any cross validation. This leads to faster computations but it does not allow computing the Q² statistics nor the confidence intervals.

- **Jackknife (LOO)**: Choose the jackknife leave one out option to recompute the model after removing each observation from the training set, one by one. This allows computing the Q² statistics as well as confidence variables on many statistics of the PLS regression such as VIP s. This option can be used if there are less than 100 observations to avoid memory problems.

- **Jackknife**: Choose the general jackknife option to recompute the model after grouping the observations and recomputing the model after removing each group from the training set, one by one. This allows computing the Q² statistics as well as confidence variables on many statistics of the PLS regression such as VIP s. This option can be used if there are less than 100 groups to avoid memory problems.

Options for PCR regression:

Standardized PCA: Activate this option to run a PCA on the correlation matrix. Inactivate this option to run a PCA on the covariance matrix (unstandardized PCA).

Filter components: You can activate one of the two following options in order to reduce the number of components used in the model:

- **Minimum %**: Activate this option and enter the minimum percentage of total variability that the selected components should represent.
- **Maximum number**: Activate this option to set the maximum number of components to take into account.

Sort components by: Choose one of the following options to determine "which criterion should be used to select the components on the basis of the "Minimum %", or of the "Maximum number":

- **Correlations with Ys**: Activate this option so that the components selection is based on the sorting down of R^2 coefficient between the dependent variable Y and the components. This option is recommended.
- **Eigenvalues**: Activate this option so that the selection of the components is based on the sorting down of the eigenvalues corresponding to the components.

Fixed intercept: Activate this option to set the intercept (or constant) of the model to a given value. Then enter the value in the corresponding field (0 by default).

Tolerance: Activate this option to allow the OLS algorithm to automatically remove the variables that would either be constant or highly correlated with other variables or group of variables (Minimum and default value is 0.0001. Maximum value allowed is 1). The higher the tolerance, the more the model tolerates collinearities between the variables.

Validation tab:

Validation: Activate this option if you want to use a sub-sample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.

- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

Quantitative: Activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

Qualitative: Activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

Observations labels: Activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Variable labels: Activate this option if the first row of the data selections (explanatory variables, observations labels) includes a header.

Missing data tab:

These options are available only for PCR and OLS regression. With PLS regression, the missing data are automatically handled by the algorithm.

Remove observations: Activate this option to remove the observations with missing data.

- **Check for each Y separately**: Choose this option to remove the observations with missing data in the select Y (dependent) variables, only if the Y of interest has a missing data.
- **Across all Ys**: Choose this option to remove the observations with missing data in the Y (dependent) variables, even if the Y of interest has no missing data.

Ignore missing data: Activate this option to ignore missing data. If missing data are present for the dependent variable, the corresponding observations will be predicted. If missing data are present for the explanatory variable(s) the corresponding observations are used to estimate the correlation matrix with pairwise deletion.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Options common to all methods:

Descriptive statistics: Activate this option to display the descriptive statistics for all the selected variables.

Correlations: Activate this option to display the correlation matrix for the quantitative variables (dependent and explanatory).

Standardized coefficients: Activate this option to display the standardized parameters of the model (also called beta coefficients).

Equation: activate this option to explicitly display the equation of the model.

Predictions and residuals: Activate this option to display the table of predictions and residuals.

Options for PLS regression:

t, u and u~ components: Activate this option to display the tables corresponding to the components. If this option is not activated the corresponding charts are not displayed.

c, w, w* and p vectors: Activate this option to display the tables corresponding to the vectors obtained from the PLS algorithm. If this option is not activated the corresponding charts are not displayed.

VIPs: Activate this option to display the table and the charts of the Variable Importance for the Projection.

Confidence intervals: Activate this option to compute the confidence intervals of the standardized coefficients. The computations involve a jackknife method.

Outliers analysis: Activate this option to display the table and the charts of the outliers analysis.

Options for PLS-DA:

t, u and u~ components: Activate this option to display the tables corresponding to the components. If this option is not activated the corresponding charts are not displayed.

c, w, w* and p vectors: Activate this option to display the tables corresponding to the vectors obtained from the PLS algorithm. If this option is not activated the corresponding charts are not displayed.

VIPs: Activate this option to display the table and the charts of the Variable Importance for the Projection.

Confidence intervals: Activate this option to compute the confidence intervals of the standardized coefficients. The computations involve a jackknife method.

Outliers analysis: Activate this option to display the table and the charts of the outliers analysis.

Confusion matrix: Activate this option to display the table showing the numbers of well- and badly-classified observations for each of the classes.

Options for PCR regression:

Factor loadings: Activate this option to display the factor loadings. The factor loadings are equal to the correlations between the principal components and the input variables if the PCA is based on the correlation matrix (standardized PCA).

Correlations Factors/Variables: Activate this option to display the correlations between the principal component and the input variables.

Factor scores: Activate this option to display the factor scores (coordinates of the observations in the new space) generated by the PCA. The scores are used in the regression step of the PCR.

Options for PCR and OLS regression:

Analysis of variance: Activate this option to display the analysis of variance table.

Adjusted predictions: Activate this option to compute and display the adjusted predictions in the predictions and residuals table.

Influence diagnostics: Activate this option to calculate and display the table that contains the influence statistics for each observation.

Charts tab:

Options common to PLS/PCR:

Regression charts: Activate this option to display the regression charts:

- **Standardized coefficients**: Activate this option to display a chart with the standardized coefficients of the model, and the corresponding confidence intervals.
- **Predictions and residuals**: Activate this option to display the following charts:

(1) Regression line: this chart is displayed only if there is one explanatory variable and if that variable is quantitative.

(2) Explanatory variable versus standardized residuals: this chart is displayed only if there is one explanatory variable and if that variable is quantitative.

(3) Dependent variable versus standardized residuals.

(4) Predictions versus observed values.

(5) Bar chart of the standardized residuals.

- **Confidence intervals**: Activate this option to display the confidence intervals on charts (1) and (4).

Correlation charts: Activate this option to display the charts involving correlations between the components and input variables. In the case of PCR, activate this option to display the correlation circle.

- **Vectors**: Activate this option to display the input variables with vectors.

Observations charts: activate this option to display the charts that allow visualizing the observations in the new space.

- **Labels**: Activate this option to display the observations labels on the charts. The number of labels can be modulated using the filtering option.

Biplots: Activate this option to display the charts where the input variables and the observations are simultaneously displayed.

- **Vectors**: Activate this option to display the input variables with vectors.
- **Labels**: Activate this option to display the observations labels on the biplots. The number of labels can be modulated using the filtering option.

Colored labels: Activate this option to display the labels with the same color as the corresponding points. If this option is not activated the labels are displayed in black.

Filter: Activate this option to modulate the number of observations displayed:

- **Random**: The observations to display are randomly selected. The "Number of observations" N to display must then be specified.
- **N first rows**: The N first observations are displayed on the chart. The "Number of observations" N to display must then be specified.
- **N last rows**: The N last observations are displayed on the chart. The "Number of observations" N to display must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to display.

Results

Descriptive statistics: the tables of descriptive statistics display for all the selected variables a set of basic statistics. For the dependent variables (colored in blue), and the quantitative explanatory variables, XLSTAT displays the number of observations, the number of observations with missing data, the number of observations with no missing data, the mean,

and the unbiased standard deviation. For the qualitative explanatory variables XLSTAT displays the name and the frequency of the categories.

Correlation matrix: this table is displayed to allow your visualizing the correlations among the explanatory variables, among the dependent variables and between both groups.

Results of the PLS regression:

The first table displays the **model quality** indexes. The quality corresponds here to the cumulated contribution of the components to the indexes:

- The **Q²cum** index measures the global contribution of the h first components to the predictive quality of the model (and of the sub-models if there are several dependent variables). The Q²cum(h) index writes:

$$Q^2cum(h) = 1 - \prod_{j=1}^{h} \frac{\sum_{k=1}^{q} PRESS_{kj}}{\sum_{k=1}^{q} SSE_{k(j-1)}}$$

The index involves the $PRESS$ statistic (that requires a cross-validation), and the Sum of Squares of Errors (SSE) for a model with one less component. The search for the maximum of the Q^2cum index is equivalent to finding the most stable model.

- The **R²Ycum** index is the sum of the coefficients of determination between the dependent variables and the h first components. It is therefore a measure of the explanatory power of the h first components for the dependent variables of the model.

- The **R²Xcum** index is the sum of the coefficients of determination between the explanatory variables and the h first components. It is therefore a measure of the explanatory power of the h first components for the explanatory variables of the model.

A bar chart is displayed to allow the visualization of the evolution of the three indexes when the number of components increases. While the R²Ycum and R²Xcum indexes necessarily increase with the number of components, this is not the case with Q²cum.

The next table corresponds to the **correlation matrix** of the explanatory and dependent variables with the t and u components. A chart displays the correlations with the t components.

The next table displays the **w vectors**, followed by the **w* vectors** and the **c vectors**, that are directly involved in the model, as it is shown in the "Description" section. If to $h = 2$ corresponds a valid model, it is shown that the projection of the x vectors on the y vectors on the **variables on the w*/c axes chart**, gives a fair idea of the sign and the relative weight of the corresponding coefficients in the model.

The next table displays the **scores** of the observations in the space of the **t** components. The corresponding chart is displayed. If some observations have been selected for the validation, they are displayed on the chart.

The next table displays the standardized **scores** of the observations in the space of the **t** components. These scores are equivalent to computing the correlations of each observation (represented by an indicator variable) with the components. This allows displaying the observations on the **correlations map** that follows where the X s, the Y s and the

observations are simultaneously displayed. An example of an interpretation of this map is available in Tenenhaus (2003).

The next table corresponds to the **scores** of the observations in the space of the **u** and then the **u~** components. The chart based on the u is displayed. If some observations have been selected for the validation, they are displayed on the chart.

The table with the **Q² quality indexes** allows visualizing how the components contribute to the explanation of the dependent variables. The table of the **cumulated Q² quality indexes** allows measuring the quality that corresponds to a space with an increasing number of dimensions.

The table of the **R² and redundancies** between the input variables (dependent and explanatory) and the components t and u allow evaluating the explanatory power of the t and u . The redundancy between an X table (n rows and p variables) and a c component is the part of the variance of X explained by c. We define it as the mean of the squares of the correlation coefficients between the variables and the component:

$$Rd(X,c) = \frac{1}{p}\sum_{j=1}^{p} R^2(x_j,c)$$

From the redundancies one can deduce the VIP s (**Variable Importance for the Projection**) that measure the importance of an explanatory variable for the building of the t components. The VIP for the j th explanatory variable and the component h is defined by:

$$VIP_{hj} = \sqrt{\frac{p}{\sum_{i=1}^{h} Rd(Y,t_i)} \sum_{i=1}^{h} {w_{ij}}^2 Rd(Y,t_i)}$$

On the VIP charts (one bar chart per component), a border line is plotted to identify the VIP s that are greater than 0.8 and above: these thresholds suggested by Wold (1995) and Ericksson (2001) allow identifying the variables that are moderately ($0.8 < VIP < 1$) or highly influential ($VIP > 1$).

The next table displays the **outliers analysis**. The DModX (distances from each observation to the model in the space of the X variables) allow identifying the outliers for the explanatory variables, while the DModY (distances from each observation to the model in the space of the Y variables) allow identifying the outliers for the dependent variables. On the corresponding charts the threshold values DCrit are also displayed to help identifying the outliers: the DMod values that are above the DCrit threshold correspond to outliers. The DCrit are computed using the threshold values classically used in box plots. The value of the DModX for the i-th observation writes:

$$DModX_i = \sqrt{\frac{n}{n-h-1}\frac{\sum_{j=1}^{p} e(X,t)^2{}_{ij}}{p-h}}$$

where the $e(X,t)_{ij}$ (i = 1 … n) are the residuals of the regression of X on the j-th component. The value of the DModY for the i-th observation writes:

$$DModY_i = \sqrt{\frac{\sum_{j=1}^{q} e(Y,t)_{ij}^2}{q-h}}$$

where q is the number of dependent variables and the $e(Y,t)_{ij}$, $(i = 1, \ldots, n)$ are the residuals of the regression of Y on the j-th component.

The next table displays the **parameters** of the models corresponding to the one or more dependent variables. It is followed by the equation corresponding to each model, if the number of explanatory variables does not exceed 20.

For each of the dependent variables a series of tables and charts is displayed.

Goodness of fit statistics: this table displays the goodness of fit statistics of the PLS regression model for each dependent variable. The definition of the statistics is as follows:

The table of the **standardized coefficients** (also named beta coefficients) allows comparing the relative weight of the variables in the model. To compute the confidence intervals, in the case of PLS regression, the classical formulae based on the normality hypotheses used in OLS regression do not apply. A bootstrap method suggested by Tenenhaus *et al.* (2004) allows estimating the confidence intervals. The greater the absolute value of a coefficient, the greater the weight of the variable in the model. When the confidence interval around the standardized coefficients includes 0, which can easily be observed on the chart, the weight of the variable in the model is not significant.

In the **predictions and residuals** table, the weight, the observed value of the dependent variable, the corresponding prediction, the residuals and the confidence intervals are displayed for each observation. Two types of confidence intervals are displayed: an interval around the mean (it corresponds to the case where the prediction is made for an infinite number of observations with a give set of values of the explanatory variables) and an interval around an individual prediction (it corresponds to the case where the prediction is made for only one observation). The second interval is always wider than the first one, as the uncertainty is of course higher. If some observations have been selected for the validation, they are displayed in this table.

The **three charts** that are displayed afterwards allow visualizing:

- the residuals versus the dependent variable,
- the distance between the predicted and observed values (for an ideal model the all the points would be on the bisecting line),
- the bar chart of the residuals.

If you have selected data to use in prediction mode, a table displays the **predictions on the new observations** and the corresponding confidence intervals.

PLS-DA specific results:

Classification functions: The classification functions can be used to determine which class an observation is to be assigned to using values taken for the various explanatory variables. These functions are linear. An observation is assigned to the class with the highest classification function $F()$.

Prior and posterior classification and scores: This table shows for each observation its membership class defined by the dependent variable, the membership class as deduced by the membership probabilities and the classification function score for each category of the dependent variable.

Confusion matrix for the estimation sample: The confusion matrix is deduced from prior and posterior classifications together with the overall percentage of well-classified observations.

Results of the PCR regression:

The PCR regression requires a Principal Component Analysis step. The first results concern the latter.

Eigenvalues: the table of the eigenvalues and the corresponding **scree plot** are displayed. The number of eigenvalues displayed is equal to the number of non null eigenvalues. If a components filtering option has been selected it is applied only before the regression step.

If the corresponding outputs options have been activated, XLSTAT displays the **factor loadings** (the coordinates of the input variables in the new space), then the correlations between the input variables and the components. The **correlations** are equal to the factor loadings if the PCA is performed on the correlation matrix. The next table displays the **factor scores** (the coordinates of the observations in the new space), and are later used for the regression step. If some observations have been selected for the validation, they are displayed in this table. A biplot is displayed if the corresponding option has been activated.

If the filtering option based on the correlations with the dependent variables has been selected, the components used in the regression step are those that have the greatest determination coefficients (R^2) with the dependent variables. The matrix of the correlation coefficients **between the components and the dependent variables** is displayed. The number of components that are kept depends on the number of eigenvalues and on the selected options ("Minimum %" or "Max components").

If the filtering option based on the eigenvalues has been selected, the components used in the regression step are those that have the greatest eigenvalues. The number of components that are kept depends on the number of eigenvalues and on the selected options ("Minimum %" or "Max components").

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.

- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.

- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).

- **R²**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ with } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

The R^2 is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer R^2 is to 1, the better is the model. The problem with the R^2 is that it does not take into account the number of variables used to fit the model.

- **Adjusted R²**: The adjusted determination coefficient for the model. The adjusted R^2 can be negative if the R^2 is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}$$

The adjusted R^2 is a correction to the R^2 which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W - p^*}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

- **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows:

$$MAPE = \frac{100}{W}\sum_{i=1}^{n} w_i \left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

- **DW**: The Durbin-Watson statistic is defined by:

$$DW = \frac{\sum_{i=2}^{n}[(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2}$$

This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- **Cp**: Mallows Cp coefficient is defined by:

$$Cp = \frac{SSE}{\hat{\sigma}} + 2p^* - W$$

where SSE is the sum of the squares of the errors for the model with p explanatory variables and $\hat{\sigma}$ is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to p*, the less the model is biased.

- **AIC**: Akaike's Information Criterion is defined by:

$$AIC = W \ln(\frac{SSE}{W}) + 2p^*$$

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz's Bayesian Criterion is defined by:

$$SBC = W \ln(\frac{SSE}{W}) + \ln(W)p^*$$

This criterion, proposed by Schwarz (1978) is similar to the AIC, and the aim is to minimize it.

- **PC**: Amemiya's Prediction Criterion is defined by:

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}$$

This criterion, proposed by Amemiya (1980) is used, like the adjusted R^2 to take account of the parsimony of the model.

- **Press**: The Press (predicted residual error sum of squares) statistic is only displayed if the corresponding option has been activated in the dialog box. It is defined by:

$$Press = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i(-i)})^2$$

where $\hat{y}_{i(-i)}$is the prediction for observation i when the latter is not used for estimating parameters. We then get:

$$\text{Press } RMSE = \sqrt{\frac{Press}{W - p^*}}$$

Press's RMSE can then be compared to the RMSE. A large difference between the two shows that the model is sensitive to the presence or absence of certain observations in the model.

The **analysis of variance** table allows evaluating how much information the explanatory variables bring to the model. In the case where the intercept of the model is not fixed by the user, the explanatory power is measured by comparing the fit of the selected model with the fit

of a basic model where the dependent variable equals its mean. When the intercept is fixed to a given value, the selected model is compared to a basic model where the dependent model equals the fixed intercept.

In the case of a PCR regression, the first table of **model parameters** corresponds to the parameters of the model based on the selected components. This table is not easy to interpret. For that reason a transformation is performed to obtain the **parameters of the model** corresponding to the input variables. The latter table is directly obtained in the case of an OLS regression. In this table you will find the estimate of the parameters, the corresponding standard error, the Student's t, the corresponding probability, as well as the confidence interval.

The **equation of the model** is then displayed to facilitate the visualization or the reuse of the model.

The table of the **standardized coefficients** (also named beta coefficients) allows comparing the relative weight of the variables in the model. The greater the absolute value of a coefficient, the greater the weight of the variable in the model. When the confidence interval around the standardized coefficients includes 0, which can easily be observed on the chart, the weight of the variable in the model is not significant.

In the **predictions and residuals** table, the weight, the value of the explanatory variable if there is only one, the observed value of the dependent variable, the corresponding prediction, the residuals and the confidence intervals and the adjusted prediction are displayed for each observation. Two types of confidence intervals are displayed: an interval around the mean (it corresponds to the case where the prediction is made for an infinite number of observations with a give set of values of the explanatory variables) and an interval around an individual prediction (it corresponds to the case where the prediction is made for only one observation). The second interval is always wider than the first one, as the uncertainty is of course higher. If some observations have been selected for the validation, they are displayed in this table.

The **charts** that follow allow visualizing the results listed above. If there is only one explanatory variable in the model, and if that variable is quantitative, then the first chart allows visualizing the observations, the regression line and the confidence intervals around the prediction. The second chart displays the standardized residuals versus the explanatory variable. The residuals should be randomly distributed around the abscissa axis. If a trend can be observed, that means there is a problem with the model.

The **three charts** that are displayed afterwards allow visualizing respectively the standardized residuals versus the dependent variable, the distance between the predicted and observed values (for an ideal model the all the points would be on the bisecting line), and the bar chart of the standardized residuals. The third chart makes it possible to quickly see if there is an unexpected number of high residuals: the normality assumption for the residuals is such that only 5% of the standardized residuals should be out of the]-2, 2[interval.

If you have selected data to use in prediction mode, a table displays the **predictions on the new observations** and the corresponding confidence intervals.

The table of **influence diagnostics** displays for each observation, its weight, the corresponding residual, the standardized residual (divided by the $RMSE$), the studentized residual, the deleted residual, the studentized deleted residual, the centered leverage, the Mahalanobis distance, the Cook's D, the CovRatio, the DFFit, the standardized DFFit, the DFBetas (one per model coefficient) and the standardized DFBetas.

Three charts are then displayed to make possible an easy identification of the observations which influence on the predictions or on the coefficients requiring a special investigation.

Examples

A tutorial on how to use PLS regression is available on the XLSTAT Help Center:

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

References

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.

Am emiya 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, MA.

Eriksson L., Johansson E., Kettaneh-Wold N. and Wold S. (2001). Multi- and Megavariate Data Analysis. Principles and Applications, Umetrics Academy, Umeå.

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.

LASSO Regression

Use this method to perform a regression when you have more variables than observations or, more universally, when there is a large number of variables.

In this section:

Description

LASSO stands for *Least Absolute Shrinkage and Selection Operator*. The LASSO regression was proposed by Robert Tibshirani in 1996. It is an estimation method that constrains its coefficients not to *explode*, unlike standard linear regression in the field of high-dimensional statistics. The high-dimensional context covers all situations where we have a very large number of variables compared to the number of individuals.

LASSO regression is one of the methods that overcome the shortcomings (instability of the estimate and unreliability of the prediction) of linear regression in a high-dimensional context. The main advantage of LASSO regression is its ability to perform variable selection, which can be valuable when there are a large number of variables.

LASSO regression

Noting Y as the vector of the quantitative dependent variable and X as the matrix of explanatory variables, the LASSO estimator $\hat{\beta}$ is the solution for the following constrained minimization problem:

$$ argmin_{\beta \in \mathbb{R}^{p}} L(\beta)=\left|Y-X\beta\right|^2 $$ subject to the constraints $\sum_{j=1}^{p} |\beta_j| \leq t$ for some $t > 0$ and where p represents the number of variables.

The Lagrangian associated with the optimization problem is written:

$$\|Y - X\beta\|^2 + 2\lambda(\sum_{j=1}^{p}|\beta_j| - t)$$

where 2λ is the Lagrange multiplier related to t by the constraint $\sum_{j=1}^{p}|\beta_j| = t$.

On the other hand, there is no explicit formula for the solution $\hat{\beta}$ for a given λ. Therefore, the optimization problem is solved using algorithms. In XLSTAT, this solution is done using the coordinate descent algorithm.

Coordinate descent algorithm.

Using the Lagrangian form, we express, for a given $\lambda > 0$, the Lasso estimator $\hat{\beta}$ as $argmin_{\beta \in \mathbb{R}^p}$ of:

$$L(\beta) = \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

The LASSO regression is based on a fundamental parameter: the λ regularization parameter. XLSTAT offers to its users to find this optimal λ parameter by cross-validation.

In a coordinate descent algorithm, the optimization of each parameter is done separately (keeping all others fixed).

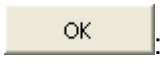
Noting that each variable X_j is centered and reduced, the j-th coordinate $\hat{\beta}_j$ of the LASSO solution is expressed as the following way:

$$ \hat\beta_j = \left\lbrace \begin{aligned} & 0 & if & \lvert R_j \rvert \leq \lambda \ & (R_j-\lambda)/n & if & R_j > \lambda \ & (R_j+\lambda)/n & if & R_j < -\lambda \end{aligned} \right. $$ with $R_j = X_j'Y - X_j' \sum_{k \neq j} X_k \hat{\beta}_k$ the j-th partial residual, X_j' the transpose of the variable X_j and n the number of observations.

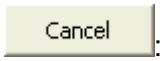
Coordinate descent algorithms use this formula to update each coordinate of the LASSO estimator until convergence of L(β) is reached.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Y / Dependent variables:

Quantitative: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

Response type: Select the type of response you have:

- **Quantitative**: If your response type contains real values, choose this type to fit a regression model.

X / Explanatory variables:

Quantitative: Select the quantitative explanatory variables in the Excel worksheet. The data selected must be numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Variable labels" option is activated, you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

Options tab:

Cross-validation: Activate this option if you want to calculate the λ parameter by cross-validation. This option allows you to run a k-fold cross-validation to obtain the optimal λ regularization parameter. This option allows you to run a k-fold cross-validation to quantify the quality of the classifier or the regression with chosen parameters. Data is partitioned into k subsamples of equal size. A single subsample is retained as the validation data to test the model, and the remaining k-1 subsamples are used as training data.

- **Number of folds**: Enter the number of folds to be constituted for the cross validation. Default value: 5.
- **Number of values tested**: Enter the number of λ values that will be tested during the cross validation. Default value: 100.

Lambda: Activate this option if you want to specify the λ regularization parameter.

Stop conditions:

- **Convergence**: Enter the maximum value of the evolution of the log of the likelihood from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.
- **Maximum time (in seconds)**: Enter the maximum time allowed for a coordinate descent. Past that time, if convergence has not been reached, the algorithm stops and returns the results obtained during the last iteration. Default value: 180 seconds.

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 5).

Validation tab:

Validation: Activate this option if you want to use a sub-sample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: Activate this option if you want to select data to use in prediction mode. If you activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

Quantitative: Select the quantitative explanatory variables in the Excel worksheet. The first row must not include variable labels.

Observations labels: Activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Variable labels: Activate this option if the first row of the data selections (variables, observations labels) includes a header.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics of the training sample for the variables selected.

- **Validation set**: Activate this option to also display descriptive statistics of the validation sample for the variables selected.
- **Prediction set**: Activate this option to also display descriptive statistics of the prediction sample for the variables selected.

Correlation matrix: Activate this option to display a view of the correlations between the various variables selected.

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

All coefficients: Activate this option to display the variables associated with zero coefficients in the results.

Charts tab:

Predictions and residuals: Activate this option to display the following charts:

- Dependent variable versus residuals.
- Predictions for the dependent variable versus residuals.
- Predictions for the dependent variable versus the dependent variable.
- Bar chart of residuals.

Variable importance: Activate this option to display the chart showing the variable importance measures.

Evolution of the MSE (Cross-validation): Activate this option to display the chart showing the evolution of the MSE according to the λ parameter.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables.

Correlation matrix: This table is displayed to give you a view of the correlations between the various variables selected.

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.
- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).
- **R²**: The determination coefficient for the model. This coefficient, which value is between 0 and 1, is defined by:

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ with } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

The R² is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer R² is to 1, the better is the model.

- **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

Model parameters: This table gives the value of each parameter after fitting it to the model

Standardized coefficients: The table of standardized coefficients (also called beta coefficients) are used, if the matrix containing the explanatory variables has not been centered, to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable.

Chart of variable importance: The importance measure for a given variable is the absolute value of its coefficient in the regression.

Chart of MSE evolution (Cross-validation): This chart shows the MSE evolution according to the λ parameter.

Predictions and residuals: This table shows, for each observation, the observed value of the dependent variable, the model's prediction and the residuals.

Charts of predictions and residuals: These charts allow you to visualize the results mentioned above.

Example

A tutorial on how to use the LASSO regression is available on the XLSTAT Help Center:

https://help.xlstat.com/6423-lasso-regression-excel

Bibliographie

Frédéric Lavancier (2020). Statistique en grande dimension.

Jerome Friedman, Trevor Hastie and Rob Tibshirani (2008). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Volume 2.

Jerome Friedman, Trevor Hastie and 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).

Ridge Regression

Use this method to perform a regression when you have more variables than observations or, more universally, when the number of variables is large.

In this section:

Description

Ridge regression, a method derived from Tikhonov regularization, was proposed by Hoerl and Kennard in 1970. It is an estimation method that constrains its coefficients not to *explode*, unlike standard linear regression in the field of high-dimensional statistics. The high-dimensional context covers all situations where we have a very large number of variables compared to the number of individuals.

Ridge regression is one of the methods that overcome the shortcomings (instability of the estimate and unreliability of the prediction) of linear regression in a high-dimensional context. Ridge regression differs from LASSO regression in that it shows greater robustness when datasets with high multicollinearity are involved.

Ridge regression

Noting Y as the vector of the quantitative dependent variable and X as the matrix of explanatory variables, the Ridge estimator $\hat{\beta}$ is the solution of the following constrained minimization problem:

$$ argmin_{\beta \in \mathbb{R}^{p}} L(\beta)=\left|Y-X\beta\right|^2 $$ subject to the constraints $\sum_{j=1}^{p} (\beta_j)^2 \leq t$ for some $t > 0$ and where p represents the number of variables.

The Lagrangian associated with the optimization problem is written:

$$\|Y - X\beta\|^2 + \lambda(\sum_{j=1}^{p}(\beta_j)^2 - t)$$

where λ is the Lagrange multiplier related to t by the constraint $\sum_{j=1}^{p}(\beta_j)^2 = t$.

Unlike LASSO regression, the Ridge estimator $\hat{\beta}$ has an explicit form:

$$\hat{\beta} = (X'X + \lambda I_p)^{-1}X'Y$$

where I_p is the identity matrix of order p.

However, in high dimension, the inversion of the matrix $X'X + \lambda I_p$ can be complicated. Therefore, the optimization problem is solved using algorithms. In XLSTAT, this solution is done using coordinate descent algorithm.

Coordinate descent algorithm

Using the Lagrangian form, we express, for a given $\lambda > 0$, the Ridge estimator $\hat{\beta}$ as $argmin_{\beta \in \mathbb{R}^p}$ of:

$$L(\beta) = \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} (\beta_j)^2$$

The Ridge regression is based on a fundamental parameter: the λ regularization parameter. XLSTAT lets users find this optimal λ parameter by cross-validation.

In coordinate descent algorithms, the optimization of each parameter is done separately (keeping all others fixed).

Thus, noting that each variable X_j is centered and reduced, the j-th coordinate $\hat{\beta}_j$ of the Ridge solution is expressed as the following way:

$$ \hat\beta_j = \frac{R_j}{n(1+\lambda)} $$ with $R_j = X_j'Y - X_j' \sum_{k \neq j} X_k \hat{\beta}_k$ the j-th partial residual, X_j' the transpose of the variable X_j and n the number of observations.

Coordinate descent algorithms use this formula to update each coordinate of the Ridge estimator until convergence of $L(\beta)$ is reached.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the calculations.

Cancel: Click this button to close the dialog box without doing any calculations.

Help: Click this button to display help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Y / Dependent variables:

Quantitative: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

Response type: Select the type of response you have:

- **Quantitative**: If your response type contains real values, choose this type to fit a regression model.

X / Explanatory variables:

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If the variable header has been selected, make sure the "Variable labels" option has been activated.

Qualitative: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, make sure the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated, you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

Options tab:

Cross-validation: Activate this option if you want to calculate the λ parameter by cross-validation. This option allows you to run a k-fold cross-validation to obtain the optimal λ regularization parameter. This option allows you to run a k-fold cross-validation to quantify the quality of the classifier or the regression with chosen parameters. Data is partitioned into k equally subsamples of equal size. A single subsample is retained as the validation data to test the model, and the remaining k-1 subsamples are used as training data.

- **Number of folds**: Enter the number of folds to be constituted for the cross validation. Default value: 5.
- **Number of values tested**: Enter the number of λ values that will be tested during the cross validation. Default value: 100.

Lambda: Activate this option if you want to specify the λ regularization parameter.

Stop conditions:

- **Convergence**: Enter the maximum value of the evolution of the log of the likelihood from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.
- **Maximum time (in seconds)**: Enter the maximum time allowed for a coordinate descent. Past that time, if convergence has not been reached, the algorithm stops and returns the results obtained during the last iteration. Default value: 180 seconds.

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 5).

Validation tab:

Validation: Activate this option if you want to use a sub-sample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.

- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: Activate this option if you want to select data to use in prediction mode. If you activate this option, you need to make sure that the prediction dataset is structured like the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

Quantitative: Activate this option to select the quantitative explanatory variables. The first row should include variable labels if the Variable labels option is activated on this page.

Qualitative: Activate this option to select the qualitative explanatory variables. The first row should include variable labels if the Variable labels option is activated on this page

Observations labels: Activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Variable labels: Activate this option if the first row of the data selections (variables, observations labels) includes a header.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics of the training sample for the variables selected.

- **Validation set**: Activate this option to also display descriptive statistics of the validation sample for the variables selected.
- **Prediction set**: Activate this option to also display descriptive statistics of the prediction sample for the variables selected.

Correlation matrix: Activate this option to display a view of the correlations between the various variables selected.

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

Charts tab:

Predictions and residuals: Activate this option to display the following charts:

- Dependent variable versus residuals.
- Predictions for the dependent variable versus residuals.
- Predictions for the dependent variable versus the dependent variable.
- Bar chart of residuals.

Variable importance: Activate this option to display the chart showing the variable importance measures.

Evolution of the MSE (Cross-validation): Activate this option to display the chart showing the evolution of the MSE according to the λ parameter.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables.

Correlation matrix: This table is displayed to give you a view of the correlations between the various variables selected.

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.
- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).

- **R²**: The determination coefficient for the model. This coefficient, which value is between 0 and 1, is defined by:

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ with } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

The R² is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer R² is to 1, the better the model.

- **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

Model parameters: This table gives the value of each parameter after fitting it to the model

Standardized coefficients: The table of standardized coefficients (also called beta coefficients) are used, if the matrix containing the explanatory variables has not been centered, to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable.

Chart of variable importance: The importance measure for a given variable is the absolute value of its coefficient in the regression.

Chart of MSE evolution (Cross-validation): This chart shows the MSE evolution according to the λ parameter.

Predictions and residuals: This table shows, for each observation, the observed value of the dependent variable, the model's prediction and the residuals.

Charts of predictions and residuals: These charts allow you to visualize the results mentioned above.

Example

A tutorial on how to use the Ridge regression is available on the XLSTAT Help Center:

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

Bibliographie

Frédéric Lavancier (2020). Statistique en grande dimension.

Jerome Friedman, Trevor Hastie and Rob Tibshirani (2008). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Volume 2.

Jerome Friedman, Trevor Hastie and 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).

Elastic net Regression

Use this method to perform a regression when you have more variables than observations or, more universally, when there is a large number of variables.

In this section:

Description

Elastic net Regression is a compromise between Ridge and LASSO regression. It is an estimation method that constrains its coefficients not to *explode*, unlike standard linear regression in the field of high-dimensional statistics. The high-dimensional context covers all situations where we have a very large number of variables compared to the number of individuals.

Elastic net regression is one of the methods that overcome the shortcomings (instability of the estimate and unreliability of the prediction) of linear regression in a high-dimensional context. The idea of this method is to take advantage of the selection qualities of the LASSO estimator, while guaranteeing a better robustness in case of multicollinearity, a property inherent to Ridge regression.

Elastic net regression

Denoting Y as the vector of the quantitative dependent variable and X as the matrix of explanatory variables, the Elastic net estimator $\hat{\beta}$ is the solution for the following constrained minimization problem:

$$ argmin_{\beta \in \mathbb{R}^{p}} L(\beta)=\left|Y-X\beta\right|^2 $$ subject to the constraints $\sum_{j=1}^{p}$ $ ((1-\alpha)\beta_j^2 + \alpha\lvert\beta_j \rvert)\leq t$ for some $t > 0$ and where α is the mixing parameter (which is between 0 and 1) and p represents the number of variables.

The Lagrangian associated with the optimization problem is written:

$$\|Y - X\beta\|^2 + \lambda(\sum_{j=1}^{p}((1-\alpha)\beta_j^2 + \alpha|\beta_j|) - t)$$

where λ is the Lagrange multiplier related to t by the constraint $\sum_{j=1}^{p}(1-\alpha)\beta_j^2 + \alpha|\beta_j| = t$.

On the other hand, there is no explicit formula for the solution $\hat{\beta}$ for a given λ. Therefore, the optimization problem is solved using algorithms. In XLSTAT, this solution is done using the coordinate descent algorithm.

Coordinate descent algorithm.

Using the Lagrangian form, we express, for a given $\lambda > 0$, the Elastic net estimator $\hat{\beta}$ as $argmin_{\beta \in \mathbb{R}^p}$ of:

$$L(\beta) = \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} ((1-\alpha)\beta_j^2 + \alpha|\beta_j|)$$

The Elastic net regression is based on two fundamental parameters: the mixing parameter α (which is between 0 and 1) and the λ regularization parameter > 0. XLSTAT offers to its users to find these optimal parameters by cross-validation.

In the coordinate descent algorithm, the optimization of each parameter is done separately (keeping all others fixed).

Taking into account that each X_j variable is centered and reduced, the j-th coordinate $\hat{\beta}_j$ of the Elastic net solution is expressed in the following way:

$$ \hat\beta_j = \left\lbrace \begin{aligned} & 0 & if & \lvert R_j \rvert \leq \lambda \alpha \\ & (R_j-\lambda \alpha)/(n(1+\lambda(1-\alpha))) & if & R_j > \lambda \alpha \\ & (R_j+\lambda \alpha)/(n(1+\lambda(1-\alpha))) & if & R_j < -\lambda \alpha \end{aligned} \right. $$ with $R_j = X_j'Y - X_j' \sum_{k \neq j} X_k \hat{\beta}_k$ the j-th partial residual, X_j' the transpose of the variable X_j and n the number of observations.

Coordinate descent algorithms use this formula to update each coordinate of the Elastic net estimator until convergence of $L(\beta)$ is reached.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results:

OK: Click this button to start the calculations.

Cancel: Click this button to close the dialog box without doing any calculations.

Help: Click this button to display help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Y / Dependent variables:

Quantitative: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each variable independantly. If a column header has been selected, check that the "Variable labels" option has been activated.

Response type: Select your type of response variable:

- **Quantitative**: If your response variable is numerical, choose this type to fit a regression model.

X / Explanatory variables:

Quantitative: Select the quantitative explanatory variables in the Excel worksheet. The selected data must be numerical. If the variable header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observation labels) includes a header.

Observation labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Variable labels" option is activated, you need to include a header in the selection. If this option is not activated, the observation labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

Options tab:

Model parameters: This option allows you to choose the method to define the model parameters. * **Cross-validation**: Activate this option if you want to calculate the model parameters by cross-validation. This option allows you to run a k-folds cross-validation to obtain the optimal λ regularization parameter and to quantify the quality of the model. The data is partitioned into k subsamples of equal size. A single subsample is retained as the validation data to test the model, and the remaining k-1 subsamples are used as training data. * **Manual entry**: Activate this option if you want to specify the model parameters.

Lambda: Activate this option if you want to calculate the λ parameter by cross validation. Otherwise, enter the value you want to assign to the λ parameter.

Alpha: Activate this option if you want to calculate the α parameter by cross validation. Otherwise, enter the value you want to assign to the α parameter.

Cross-validation parameters: * **Number of folds**: Enter the number of folds to be constituted for the cross validation. Default value: 5. * **Number of values tested**: Enter the number of values for each parameter that will be tested during the cross validation. Default value: 100.

Stop conditions:

- **Convergence**: Enter the maximum value of the evolution of the log of the likelihood from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.
- **Maximum time (in seconds)**: Enter the maximum time allowed for a coordinate descent. Past that time, if convergence has not been reached, the algorithm stops and returns the results obtained during the last iteration. Default value: 180 seconds.

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 5).

Validation tab:

Validation: Activate this option if you want to use a sub-sample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.

- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: Activate this option if you want to select data to use in prediction mode. If you activate this option, you need to make sure that the prediction dataset is structured exactly like the estimation dataset: the same variables to be selected in the same order. On the other hand, variable labels must not be selected.

Quantitative: Select the quantitative explanatory variables in the Excel worksheet. The first row must not include variable labels.

Observations labels: Activate this option if observation labels are available. Then select the corresponding data. If this option is not activated, the observation labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Variable labels: Activate this option if the first row of the data selections (variables, observations labels) includes a header.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbour**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbour of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics of the training sample for the selected variables.

- **Validation set**: Activate this option to also display descriptive statistics of the validation sample for the selected variables.
- **Prediction set**: Activate this option to also display descriptive statistics of the prediction sample for the selected variables.

Correlation matrix: Activate this option to display a view of the correlations between the various selected variables.

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) of the model to be displayed.

Predictions and residuals: Activate this option to display the predictions and residuals for all of the observations.

Evolution of the MSE (Cross-validation): Activate this option to display the evolution of the MSE depending on the model parameters.

All coefficients: Activate this option to display the variables associated with zero coefficients in the results.

Charts tab:

Predictions and residuals: Activate this option to display the following charts:

- Dependent variable versus residuals.
- Predictions for the dependent variable versus residuals.
- Predictions for the dependent variable versus the dependent variable.
- Bar chart of residuals.

Variable importance: Activate this option to display the chart showing the variable importance values.

Evolution of the MSE (Cross-validation): Activate this option to display the chart showing the evolution of the MSE depending on the model parameters.

Results

Descriptive statistics: The table of descriptive statistics shows basic statistics for all of the selected variables. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables.

Correlation matrix: This table is displayed to give you a view of the correlations between the various selected variables.

Goodness of fit statistics: The statistics related to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.
- **DF**: The number of degrees of freedom for the chosen model.
- **R²**: The determination coefficient for the model. This coefficient, which value is between 0 and 1, is defined by:

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ with } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

The R² is interpreted as the proportion of the variability of the dependent variable explained by the model. The closer the R² is to 1, the better is the model.

- **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

Model parameters: This table gives the value of each parameter after fitting it to the model

Standardized coefficients: If the matrix containing the explanatory variables has not been centered, the standardized coefficients (also called beta coefficients) are used in order to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable.

Chart of variable importance: The importance measure for a given variable is the absolute value of its coefficient in the regression.

Evolution of the MSE (Cross-validation): This table shows the evolution of the MSE according to the model parameters.

Chart of MSE evolution (Cross-validation): This chart shows the evolution of the MSE according to the model parameters.

Predictions and residuals: This table shows, for each observation, the observed value of the dependent variable, the prediction of the model and the residuals.

Charts of predictions and residuals: These charts enable you to visualize the results mentioned above.

Example

A tutorial on how to use the Elastic net regression is available on the XLSTAT Help Center:

https://help.xlstat.com/6788-elastic-net-regression-excel

Bibliographie

Frédéric Lavancier (2020). Statistique en grande dimension.

Jerome Friedman, Trevor Hastie and Rob Tibshirani (2008). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Volume 2.

Jerome Friedman, Trevor Hastie and 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

Fuzzy k-means clustering

Use Fuzzy k-means clustering to create homogeneous groups of objects (clusters) described by a set of quantitative variables. If the dataset contains multiple groups with unclear borders (for example, if the groups are too close), it is possible to introduce a coefficient of fuzziness that allows each observation to be linked to each group with a probability of membership, this is soft or fuzzy clustering.

In this section :

Description

Principle of the k-means method

K-means clustering was introduced by McQueen in 1967. Other similar algorithms had been developed by Forgy (1965) (moving centers). This algorithm uses group barycenters which are updated at each iteration. Each partition is characterized by an objective function Q (or clustering criterion).The algorithm computes each iteration only if the difference between the two consecutive Q ($dQ = Q_{i-1} - Q_i$) is lower than a threshold defined by the user. The result of the algorithm strongly depends on the starting point so by multiplying the starting points and the repetitions, several solutions may be explored. This reduces the probabilities to converge into a local optimum. The disadvantage of this method is that it does not give a consistent number of clusters.

K-means algorithm allows using several dissimilarity indexes, the euclidean distance being the most used. But in cases where it is needed to reduce scaling effect like in text analysis, some distances grant better understanding of the structure. This is the case of the cosine dissimilarity which characterizes the spherical k-means algorithm.

The spherical k-means is derived from the classical k-means but uses the cosine dissimilarity between two observations which characterizes the angular distance without taking account of the sizes of the observations. This is particularly useful in text mining because two documents with the same relative amount of words are close regardless of the size of the documents.
Furthermore, the method allows the algorithm to be very optimized thanks to sparse matrix structures. This method was introduced by Dhillon in 2001.

Data matrix extracted from textual analysis (Term-Document Matrix) usually contains few positive values. This type of matrix is called "Sparse Matrix" because of the large amount of

zero values (at least 90% across all the observations). The structure of these matrices can be exploited using specific memory containers allowing to optimize both memory occupation and computation speed. XLSTAT uses a specific sparse matrix container called "Row Compressed Matrix" which only preserves non-zero values and their coordinates.

Principle of the fuzzy k-means method

The fuzzy clustering allows to create clusters with unclear borders either because they are too close or even overlap each other. This method was introduced in 1973 by Dunn and Bezdek in 1981. It can highlight sub-clusters and even predict an estimation of the right number of clusters by processing the data with a high number of clusters.

Fuzzy k-means is a generalization of the classical k-means where each observation is associated to each cluster with a probability $\mu_{i,j}$ .A starting point is chosen by associating the k centers to k observations (randomly or not). Then the distances between the observations and the centers are computed. Next, the membership probability $\mu_{i,j}$ is computed for each observation i and each center j, as follows :

$$\mu_{i,j} = \frac{\frac{1}{w_i d(X_i,C_j)}^{\frac{1}{m-1}}}{\sum_{l=1}^{k} \frac{1}{w_i d(X_i,C_l)}^{\frac{1}{m-1}}}$$

Then each center C_j is updated using the membership probability and the fuzzy coefficient m :

$$C_j = \frac{\sum_{i=1}^{N}\sum_{j=1}^{k} w_i \mu_{i,j}^m X_i}{\sum_{i=1}^{N}\sum_{j=1}^{k} w_i \mu_{i,j}^m}$$

The fuzzy coefficient should be higher than 1. The higher the coefficient is, the fuzzier the borders of the clusters are (WARNING : The fuzzy coefficient has to be carefully chosen. If this coefficient is too high, the membership probabilities may be equals for all clusters and the partition is then false). K-means is a specific case of the fuzzy k-means where $m = 1$ and $\mu_{i,j} = 1$ if $j = min_j D(X_i, C_j)$, else $\mu_{i,j} = 0$.

Dissimilarity Index and Clustering Criterion :

Several dissimilarity indices may be used to reach a solution. XLSTAT offers three distances detailed by Chuanren Liu, Tianming Huy, Yong Gez and Hui Xiongx :

- **Cosine Dissimilarity** : The cosine dissimilarity is the distance which characterizes the spherical k-means and is based on the cosine of the angle between two observations. The wider the angle, the more the cosine dissimilarity will be close to 1, with 1 being an angle of 90° meaning no variables are shared between the observations.

$$D_{cosine}(A,B) = 1 - cos(A,B) = 1 - \frac{AB^T}{||A||||B||}$$

In case of textual analysis where the scaling effect has to be small, the cosine dissimilarity is recommended.

- **Jaccard Dissimilarity**: This distance is based on the extended Jaccard index. The basic Jaccard index computes the binary intersection domain between two binary vectors over

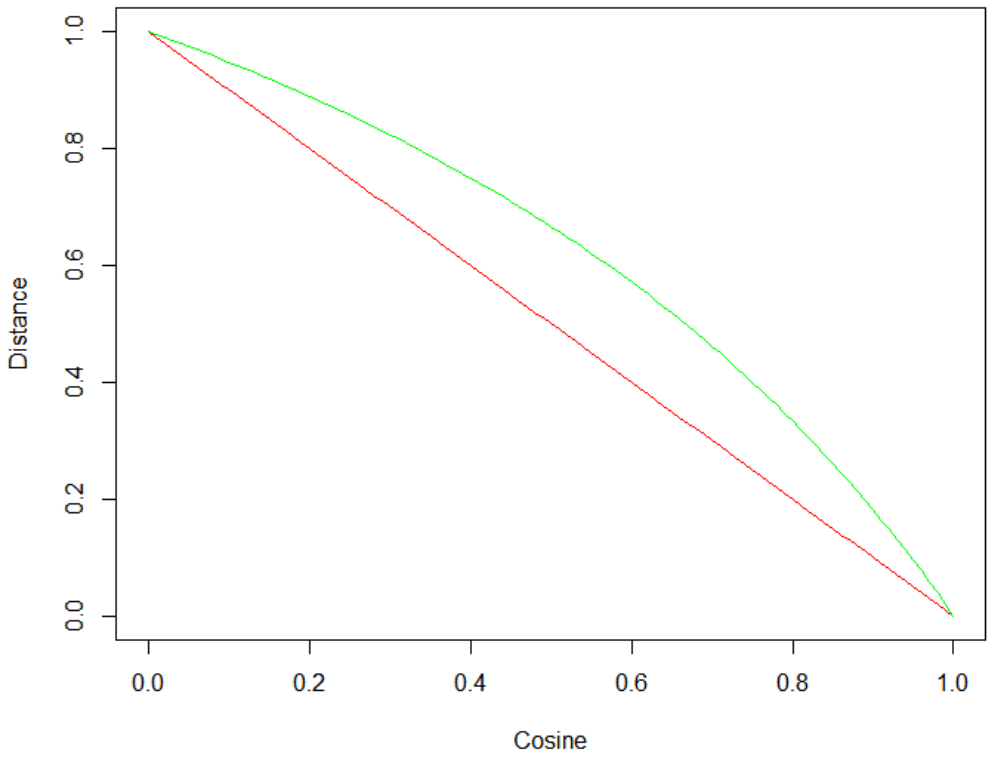
the binary union between these observations.

$$D_{Jacc}(A,B) = 1 - Jacc(A,B) = 1 - \frac{A \cap B}{A \cup B}$$

The extended jaccard index does the same thing but considers the values of the vectors as weights. In order to optimize the computation, we base the extended Jaccard index on the cosine similarity.

$$D_{Jacc_{Extended}}(A,B) = 1 - Jacc_{Extended}(A,B) = 1 - \frac{cos(A,B)}{||A||^2 + ||B||^2 - cos(A,B)}$$

Compared to the cosine dissimilarity, the extended Jaccard index has a higher sensitivity concerning close observations, which is useful for dense datasets with several close clusters.



Comparison between the cosine distance (red) and the extended Jaccard distance (green).

- **Euclidean distance**: The euclidean distance is commonly used in statistical analysis and produces, in most cases, descent results. But, keep in mind that is due to the optimization process,concerning sparse data the other two distances are recommended.

The clustering criterion Q (or objective function) is computed depending on the choice of clustering distance : for the euclidean distance three choices are available ($Trace(W)$, $Determinant(W)$, $Wilks'Lambda$) while for the Jaccard index we use the $Trace(W)$ and for Cosine dissimilarity it is the sum of distances between each observations and centers weighted by μ and m. Each criterion is described later in the document.

In the spherical case :

$$Q = \sum_{i=1}^{N} \sum_{j=1}^{k} \mu_{i,j}^{m} D_{cosine}(X_i, C_j)$$

The above clustering are described as following:

- **Trace(W)**: The W trace is the most traditional criterion. Minimizing the W trace for a given number of clusters amounts to minimizing the total within-class variance, in other words minimizing the heterogeneity of the clusters. This criterion is sensitive to effects of scaling. In order to avoid giving more weight to certain variables than others, the data must be normalized beforehand. Moreover, this criterion tends to produce clusters of the same size.

In the hard case :

$$W = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{k} \mu_{i,j} (X_i - C_j)(X_i - C_j)^T$$

Then the criterion is :

$$Q = \sum_{ii=1}^{D} W(ii, ii)$$

In the soft case, The W matrix is composed on k matrix F_j :

$$F_j = \frac{1}{\sum_{i=1}^{N} \mu_{i,j}} \sum_{i=1}^{N} \mu_{i,j} (X_i - C_j)(X_i - C_j)^T$$

Then the criterion is :

$$Q = \sum_{j=1}^{k} Trace(F_j) = \sum_{j=1}^{k} \sum_{ii=1}^{D} F_j(ii, ii)$$

- **Determinant(W)**: The determinant of W, *pooled within covariance matrix*, is a criterion considerably less sensitive to effects of scale than the W trace criterion. Furthermore, cluster sizes may be less homogeneous than with the trace criterion.

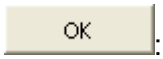
In the soft case, the criterion is :

$$Q = \sum_{j=1}^{k} \sqrt{Determinant(F_j)}$$

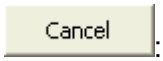
- **Wilks lambda**: The results given by minimizing this criterion are identical to that given by the determinant of W. Wilks' lambda criterion corresponds to the division of $determinant(W)$ by $determinant(T)$ where T is the total variance matrix. Dividing by the determinant of T always gives a criterion between 0 and 1.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Observations/variables table: Select a table comprising N observations described by P descriptors. If column headers have been selected, make sure the "Variable labels" option has been activated.

Observation weights: Activate this option if the rows are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, make sure the "Column labels" option is activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (Observations/variables table, Variable labels, Observation weights, Observation weights) contains a label.

Observation labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Column labels" option is activated you need to include a header in the selection. If this option is not activated, the observation labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Dissimilarity Index: Choose a distance among the three available on XLSTAT. The cosine dissimilarity is recommended in order to analyze textual data. The Jaccard index is recommended for datasets which require a fine analysis. The Euclidean distance is the most used distance for k-means and is applied in most cases.

Clustering criterion: Choose a criterion among the three available on XLSTAT. $Trace(W)$ is the within-class sum of squares. The Wilks' Lambda allows an absolute representation of the

partition. The $determinant(W)$ is a criterion that is less sensible to scaling effect than $trace(W)$.

Options tab:

Cluster rows: Activate this option if you want to create clusters of objects in rows described by descriptors in columns.

Cluster columns: Activate this option if you want to create clusters of objects in columns described by descriptors in rows.

Standardize :

- **Center** : This option computes the mean of each variable and subtracts it from each variable. This option allows the algorithm to take into account the relative statistical properties of each variable. This transformation will only be applied on variables.
- **Reduce** : This option computes the standard deviation for each variable and divides each variable by this deviation. This reduces the scale effect of the variables. This transformation will only be applied on variables.

Type of clustering :

- **Hard**: Choose this option to compute hard k-means algorithm.
- **Fuzzy**: Choose this option to compute fuzzy k-means algorithm. The default coefficient of fuzziness is 1,05.

Initial Partition : Use these options to choose the way the first partition is chosen, in other words, the way observations are assigned to clusters in the first iteration of the clustering algorithm.

- **Random**: Observations are assigned to clusters randomly.
- **Defined by centers**: Select the k centers corresponding to the k clusters. The number of rows must be equal to the number of clusters and the number of columns equal to the number of columns in the data table. If the "Column labels" option is activated you need to include a header in the selection.
- **Defined by memberships**: The observations are affected to the clusters following an index variable defined by the user (for example : 2,2,3,1,4). In this case, select a column vector containing the same number of rows as the number of observations.
- **K++**: This option lets you define centers according to k-means++ algorithm introduced by Rafail Ostrovsky, Yuval Rabani, Leonard Schulman and Chaitanya Swamy in 2006. The first center is chosen randomly among the observations. The next is chosen among the observations depending on the distance between the observation and the center. The further the observation is from the center the higher the probability it will be chosen. The $k-2$ remaining centers are chosen according to the same method. This method allows you to start with centers chosen evenly in the dataset which generally increase the quality of the partition and the speed at which the algorithm reaches the solution. But this

algorithm take times to compute and with large and complex datasets (with a lot of centers), it is recommended to use K|| algorithm.

- **K||**: This option lets you define centers according to K|| or to the "Scalable K-means" algorithm introduced by Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar and Sergei Vassilvitskii in 2012. It is derived from the K++ algorithm and lets you choose the centers with parallelization. Like K++, The first center is chosen randomly among the observations, but at each iteration of the algorithm, $k-2$ observations are chosen randomly and independently with the same method as K++. After a number of runs, depending on the size of the dataset, the X centers obtained are then reclustered into k centers using K++. This algorithm has the advantage of being much faster than K++ for two reasons : the first step is mainly decreasing the amount of relevant observations to be processed by K++ and the independent choice of the centers for each iteration allows for a parallel implementation of the first step.

Number of classes: Enter the number of clusters to be created by the algorithm. You can let the number of clusters vary between two bounds, except when "Defined by centers" or "Defined by memberships" are selected.

Number of runs: Enter the number of times you want to run this algorithm. With some parameters being random, like the selection of initial centers, running the algorithm several times helps reaching a global solution of the dataset. Using this option, only the best run is kept.

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the k-means algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 50. If the number of iteration is 0, then the algorithm will run until the convergence reaches the threshold of convergence.

- **Convergence**: Enter the minimum value of evolution for the chosen criterion from one iteration to another, after that, we consider that the algorithm converged. Default value: 0.00001. If the convergence is 0 then the algorithm will run until it reaches the maximum number of iterations.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Replace values by 0 : Activate this option to replace all missing values by 0.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

This tab is split into two parts : global results which concern or summarize the clusterings and the results by clustering.

Global results

- **Summary table**: Activate this option to display the summary of each clustering. This includes the number of clusters and iterations, the clustering criterion, the within-class and between-class sum of squares and the mean width of the silhouette.
- **Descriptive statistics**: Activate this option to display descriptive statistics for the selected variables.
- **Cluster size**: Activate this option to display the number of observations for each cluster.

Results by class

- **Centers**: Activate this option to display the cluster coordinates.
- **Central objects**: Activate this option to display the coordinates of the nearest observation to the centroid for each class.
- **Cluster Summary**: Activate this option to display the characteristics of each cluster in this partition
  (within-class variance, mean, maximum and minimum distances from the cluster center) and all the observations in the clusters.
- **Most present variables**: Activate this option to display the most present variables of each cluster. The default number of words displayed is 10.
- **Memberships**: Activate this option to display the cluster associated with each observation and the distance between these two.
- **Membership probabilities**: Activate this option to display the membership probabilities $\mu_{i,j}$ for each observation (Only available with Fuzzy clustering)

Charts tab:

Evolution of the criterion : If you choose to do the clustering between two numbers of clusters, XLSTAT displays the criterion for each partition. The higher the number of clusters and the lower this criterion will be. If there is no particular structure in the dataset, the criterion will decrease steadily but if there is a structure inside the dataset, an elbow might appear on the chart at the right number of clusters.

Profile plot: Activate this option to display a plot that allows you to compare the means of the different classes that have been created.

Cluster Size: This chart represents the number of observations of each cluster.

Silhouette: Activate this option to plot the silhouette of the partition. For each observation, a fitness coefficient between -1 and 1 will be computed, 1 being perfect fit and negative values

being bad partition. All these fitness coefficients form the silhouette of the partition. The fitness coefficients are computed as follow :

$$Fit(i) = \frac{b(i) - a(i)}{max(a(i), b(i)}$$

$a(i)$ being the average distance between the observation i and all other data within the same cluster and $b(i)$ the lowest average distance between i and all observations in any other cluster .

Condensed Silhouette: Activate this option to only display a limited amount of fitness coefficients. This greatly increases the speed of display in case of large datasets. This option will only be taken into account if the dataset contains more than 500 observations.

Example

A tutorial on k-means clustering is available on the XLSTAT Help Center:

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

References

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

Classification and regression trees

Classification and regression trees are methods that deliver models that meet both explanatory and predictive goals. Two of the strengths of this method are on the one hand the simple graphical representation by trees, and on the other hand the compact format of the natural language rules. We distinguish the following two cases where these modeling techniques should be used:

- Use classification trees to explain and predict the belonging of objects (observations, individuals) to a class, on the basis of explanatory quantitative and/or qualitative variables.
- Use regression tree to build an explanatory and predictive model for a dependent quantitative variable based on explanatory quantitative and/or qualitative variables.

Note: Sometimes the term segmentation tree or decision tree is employed when talking about the above mentioned models.

In this section:

Description

Classification and regression tree analysis was proposed in different ways. AID trees (Automatic Interaction Detection) have been developed by Morgan and Sonquist (1963). CHAID (CHi-square Automatic Interaction Detection) was proposed by Kass (1980) and later enriched by Biggs (Biggs et al, 1991) when he introduced the exhaustive CHAID procedure. The name of the Classification And Regression Trees (C&RT) methods, comes from the title of the book of Breiman (1984). The QUEST (QUick, Efficient, Statistical Tree) method is more recent (Loh and Shih, 1997).

These explanatory and predictive methods can be deployed when one needs to:

- Build a rules-based model to explain a phenomenon recorded through qualitative or quantitative dependent variables, while identifying the most important explanatory variables.
- Identify groups created by the rules.
- Predict the value of the dependent variable for a new observation.

CHAID, CART and QUEST

XLSTAT offers the choice between four different methods of classification and regression tree analysis: CHAID, exhaustive CHAID, C&RT and Quest. In most cases CHAID and exhaustive CHAID give the best results. In some cases the two other methods can be of interest. CHAID is the only algorithm implemented in XLSTAT that can lead to non binary tree.

CART

The procedure involves partitioning observations by creating the most homogeneous possible groups of observations from the perspective of the variable to predict. Several iterations are necessary : at each iteration we divide the observations into $k = 2$ classes to explain the dependent variable. The first division is obtained by choosing the explanatory variable which will provide the best separation of the observations on the basis of a quality measure. That division defines subpopulations ("nodes") of the tree. The process is repeated for each subpopulation until no further separation is possible. We then obtain terminal nodes called "leaves" of the tree. Each leaf is characterized by a specific path through the tree, which is called a "rule". The set of rules for all leaves define the model.

Quality measure:

- In the case of regression, the dependent variable is quantitative. In order to obtain the optimal split, we try to minimize at each node the variance of the child nodes. (t_L and t_R). The variance of a node t is defined by:

$$\sum_{X_i \in t} (Y_i - \overline{y}(t))^2$$

Where Y_i is the value of the dependent variable associated to observation i and where $\overline{y}(t)$ is the average of the outputs associated to node t.

- In the case of classification, the dependent variable Y is qualitative with J categories. The quality measure used to split a node are in this case the Gini impurity index, the information gain (entropy) or Twoing criterion.

For a node t these measures are defined such as:

$$GINI : i(t) = 1 - \sum_J p^2(j|t)$$

$$ENTROPY : i(t) = -\sum_J p(j|t) * \log\left[p(j|t)\right]$$

$$TWOING : i(t) = \frac{p_L * p_R}{4} \left[\sum_J p^2(j|t_L) - p^2(j|t_R)\right]^2$$

with $p(j|t)$ the probability of having the modality j of Y knowing that we are in the node t, t_L and t_R resp. left and right childs nodes, p_L et p_R the probability for an observation to belong resp. to the childs nodes left and right.

In the case of a quantitative explanatory variable, all the possible binary partitions are tested, so we have an infinity of possible tests. Nevertheless, the size n of the learning sample L_n being fixed, we have at most n distinct values for a quantitative variable, therefore at most $n-1$ associated binary questions.

For a qualitative explanatory variable, each grouping in two groups of k modalities is tested (i.e. $2^k - 1$ possibilities).

After each generation of a new subnode, the stop criteria are checked, and if none of the conditions are fulfilled, the node will be considered as a parent node, and the process is iterated.

Stop criteria:

- Pure node: The node contains only observations of one category or one value of the dependent variable.
- Variance equals zero: The variance of the dependent variable associated to observations of a node is null.
- No partitioning allows to improve the quality measure.
- Maximum tree depth: The level of the node has reached the user defined maximum tree depth.
- Minimum size for a parent-node: The node contains fewer observations than the user defined minimum size for a parent-node.
- Minimum size for a son-node: After splitting this node, there is at least one sub-node which size is smaller than the user defined minimum size for a child-node.
- Complexity parameter (CP): The construction of a tree does not continue unless the overall impurity is reduced by at least a factor CP. That value must be less than 1.

CHAID and exhaustive CHAID

Both CHAID and exhaustive CHAID algorithms consist of three steps: merging, splitting and stopping. A tree is grown by repeatedly using these three steps on each node starting from the root node.

The following algorithm only accepts nominal or ordinal categorical explanatory variables. When predictors are continuous, they are transformed into ordinal predictors with $K \leq 10$ categories before using the following algorithm.

- **Processing quantitatives explanatory variables:**

Given a quantitative variable X and $a_1, a_2, \ldots, a_{K-1}$ (in ascending order), an observation x_i de X is mapped into category $C(x)$ as follows:

$$C\left(x\right) = \begin{cases} 1 & x \leq a_1 \\ \mathrm{k}+1 & a_k < x \leq a_{K-1} \\ \mathrm{K} & a_{K-1} < x \end{cases}$$

If K is the desired number of categories, the break points are computed as follows:

Calculate the rank of $x_{(i)}$. Observations weights are incorporated when calculating the ranks. If there are ties, the average rank is used. Denote the rank and the corresponding values in ascending order as

$$\left\{r_{(i)}, x_{(i)}\right\}_{i=1}^{n}.$$

For each category $k = 0 \;\; a \;\; (K-1)$, $I_k = \left\{i : \left\lfloor r(i) * \frac{K}{N+1} \right\rfloor = k\right\}$ where $\lfloor x \rfloor$ is the floor integer of x. If the group I_k is non empty, $i_k = max\left\{i :\in I_k\right\}$. The break points are set equal to the x values corresponding to the i_k, excluding the largest.

- **Merging:** For all explanatory variables X_i, the process try to merge similar categories. Each final category of X_i will result in one child node if X_i is used to split the node. The merging step also calculates the p-value that is to be used in the splitting step. several steps are needed:
- **Chaid :**
- If X_i has 2 categories, go to **step 7**.
- Else, find the pair of categories of X_i that is most similar. The most similar pair is the pair whose test statistic gives the largest p-value with respect to the dependent variable Y. The p-values are compute using the Pearson's Chi-square test or the maximum likelihood ratio in classification case (Y qualitative) or by doing an ANOVA F test in regression case (Y quantitative).
- For the pair having the largest p-value (α), if $\alpha > \alpha_{merge}$, α_{merge} being a user-specified merge threshold, this pair is merged. If it does not, then go to **step 6**.
- (Optional) If the newly formed category consists of more than 2 original categories, then we find the best binary split within the compound category. Here the best split is the one which p-value is the smallest (α_{min}). Perform this binary split if $\alpha_{min} \leq \alpha_{spli-merge}$, $\alpha_{spli-merge}$ being the split threshold defined by the user.
- Go to **step 2**.
- (Optional) Any category having too few observations compared to a user-specified minimum child size is merged with the most similar other category(having largest p-value).
- (Optional) The p-values are adjusted by applying Bonferroni adjustments.
- **Exhaustive CHAID:**

In exhaustive CHAID the merging step use an exhaustive research method to merge any similar pair until only a single pair remain.

1. Set $j = 0$, we calculate the p-value based on the set of categories of X_i at this time. we note that P_j, $P_j = P_0$.
2. We find the allowable pair of categories of X that is, most similar. (pair whose test statistic gives the largest p-value with respect to the dependent variable Y).
3. We merge the pair that gives the largest p-value.
4. (Optional) If the new category just formed contains more than 2 original categories, we search a binary split of this new category that gives the smallest p-value. If this p-value is larger than the one in forming the category got by merging in the previous step, perform the binary split on that new category.
5. $j = j + 1$, we calculate the p-value P_j based on the set of categories of X_i at this time.
6. Repeat 2 to 5 until only two categories remain. Then among all the indices j, we find the set of categories such that P_j is the smallest.
7. (Optional) Any category having too few observations compared to a user-specified minimum child size is merged with the most similar other category (having largest p-value).
8. (Optional) The p-values are adjusted by applying Bonferroni adjustments.

- **Splitting:** Starting with the root node that contains all the objects, the best split variable is the one for which the p-value (or adjusted p-value) is the lowest. The split is performed if the p-value is lower than the user defined threshold. Else, we dont split the node.
- **Stopping:** For every newly created sub-node the stop criteria are checked. If none of the criteria are met, the node is treated as a parent node. The following are the stop criteria:
- Pure node: The node contains only objects of one category or one value of the dependent variable.
- Maximum tree depth: The level of the node has reached the user defined maximum tree depth.
- Minimum size for a parent-node: The node contains fewer objects than the user defined minimum size for a parent-node.
- Minimum size for a child-node: After splitting this node, there is at least one sub-node which size is smaller than the used defined minimum size for a son-node.

QUEST

This method can only be applied to qualitative dependent variables (classification). We carries out a splitting using two separate sub-steps. First, we look for the best splitting variable among the explanatory variables; second, the split point for the split variable is calculated:

- **Selection of the split variable:** for a quantitative explanatory variable, an ANOVA F-test is carried out to compare the mean values of each explanatory variable for the different categories of the qualitative dependent variable Y. In the case of a qualitative explanatory variable a Chi-square test is performed for each explanatory variable. We define X^* as the explanatory variable for which the p-values is the smallest. If the p-value corresponding to X^* is smaller than $\frac{\alpha}{p}$, where α is the user defined threshold and p is the number of explanatory variables, then X^* is chosen as the split variable. In the case where no X^* is found, Levene's F statistic is calculated for all the quantitative explanatory variables. We define by X^{**} the explanatory variable corresponding to the smallest p-value. If the p-value of X^{**} is smaller than $\frac{\alpha}{p+pX}$, pX being the number of quantitative explanatory variables, then X^{**} is chosen as the split variable. In the case where no X^{**} is found, the node is not split.

- **Choice of the split point:** in the case of a qualitative explanatory variable, the latter variable is first transformed into a qualitative variable X'. The detailed description of the transformation can be found in Loh and Shih (1997). In the case of quantitative variable, similar classes of Y are first grouped together by a k-means clustering of the mean values of X_i until obtaining two groups of classes. Then, a discriminant analysis using a quadratic model is done on these two groups of classes, in order to determine the optimal split point for that variable.

Stop conditions: For every newly created sub-node the stop criteria are checked. If none of the criteria are met, the node is treated in the same way as the root node.

the prior probability distribution for the dependent variable is needed to run the algorithm. XLSTAT gives you two options to compute it automaticaly : compute the priors using the classes repartition in the learning sample or assume that all categories of the dependent variable have the same repartition (equiprobability).

- **Stop conditions:** for every newly created sub-node the stop criteria are checked. If none of the criteria are met, the node is treated in the same way as the root node.

- Pure node: the node does only contain objects of one class or one value of the dependent variable.

- Maximum tree depth: the level of the node has reached the user defined maximum tree depth.

- Minimal parent node size: the node contains fewer objects than the user defined minimal parent node size.

- Minimal child node size: after splitting this node, a sub node would exist that size would be smaller than the used defined minimal son node size.

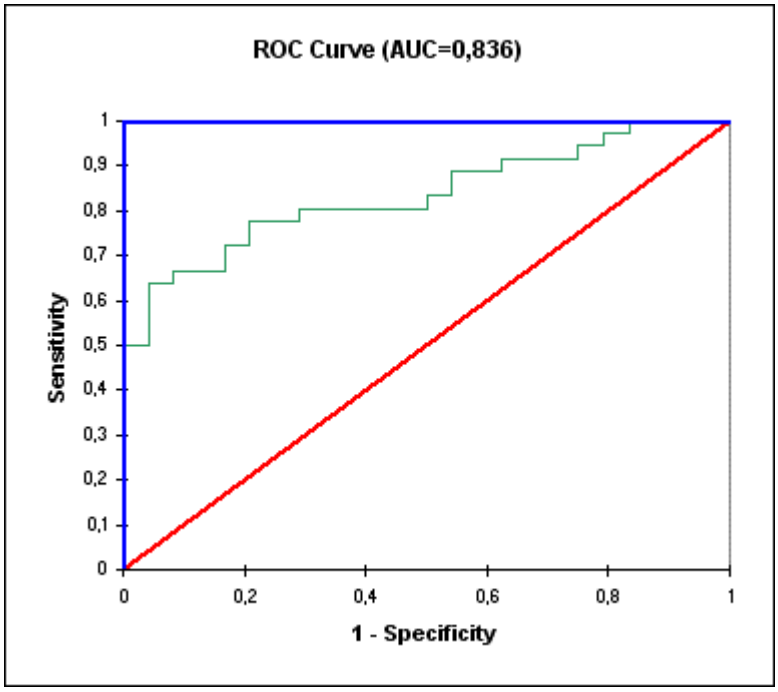
Classification table and ROC curve

Among the numerous results provided, XLSTAT can display the classification table (also called confusion matrix) used to calculate the percentage of well- classified observations. When only two classes are present in the dependent variable, the ROC curve may also be displayed.

The ROC curve (*Receiver Operating Characteristics*) displays the performance of a model and enables a comparison to be made with other models. The terms used come from signal detection theory.

The proportion of well-classified positive events is called the *sensitivity.* The *specificity* is the proportion of well-classified negative events. If you vary the threshold probability from which an event is to be considered positive, the sensitivity and specificity will also vary. The curve of points (1-specificity, sensitivity) is the ROC curve.

Let's consider a binary dependent variable which indicates, for example, if a customer has responded favorably to a mail shot. In the diagram below, the blue curve corresponds to an ideal case where the n% of people responding favorably corresponds to the n% highest probabilities. The green curve corresponds to a well-discriminating model. The red curve (first bisector) corresponds to what is obtained with a random Bernoulli model with a response probability equal to that observed in the sample studied. A model close to the red curve is therefore inefficient since it is no better than random generation. A model below this curve would be disastrous since it would be less even than random.



The area under the curve (or *AUC*) is a synthetic index calculated for ROC curves. The AUC corresponds to the probability such that a positive event has a higher probability given to it by the model than a negative event. For an ideal model, AUC=1 and for a random model, AUC = 0.5. A model is usually considered good when the AUC value is greater than 0.7. A well-discriminating model must have an AUC of between 0.87 and 0.9. A model with an AUC greater than 0.9 is excellent.

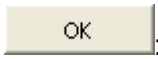
Classification and regression trees apply to quantitative and qualitative dependent variables. In the case of a Discriminant analysis or logistic regression, only qualitative dependent variables can be used. In the case of a qualitative depending variable with only two categories, the user will be able to compare the performances of both methods by using ROC curves.

Lastly, it is suggested to validate the model using a validation sample wherever possible. XLSTAT has several options for automatically generating a validation sample.

- **Lift curve** : The Lift curve is the curve that represents the Lift value as a function of the percentage of the population. Lift is the ratio between the proportion of true positives and the proportion of positive predictions. A Lift of 1 means that there is no gain over an algorithm that makes random predictions. Usually, the higher the Lift, the better the model.
- **Cumulative gain curve** : The gain curve represents the sensitivity, or recall, as a function of the percentage of the total population. It allows us to see which portion of the data concentrates the maximum number of positive events.

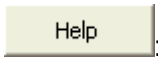
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: click this button to start the calculations.

: click this button to close the dialog box without doing any calculations.



: click this button to display help.


: click this button to reload the default options.


: click this button to delete the data selections.


: click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Dependent variables: select the dependent variable you want to model. If a column header has been selected, check that the "Variable labels" option has been activated.

Response type: Confirm the type of dependent variable you have selected:

- **Quantitative**: Choose this option if the selected dependent variable is quantitative.

- **Qualitative**: Choose this option if the selected dependent variable is qualitative.

X / Explanatory variables:

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If a variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If a variable header has been selected, check that the "Variable labels" option has been activated.

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2, …).

Method: Choose the method to be used. CHAID, exhaustive CHAID, C&RT and Quest are possible choices. In the case of Quest, the "Response type" is automatically changed to qualitative.

Measure: In the case of the CHAID or exhaustive CHAID methods with a qualitative response type, the user can choose between the Pearson's Chi-square and the likelihood ratio measures. In the case of the CART method together with a qualitative response type, the user can choose between the Gini, information gain and Twoing measures.

Options tab:

- **General** tab:

Tree parameters : Choose the way to define the parameters, **Automatic**, **manually** or using a **Grid** search.

- **Enter manually:** Enter the value of each parameter.
- **Automatic:** Select the parameters you want XLSTAT to estimate automatically. The algorithm will randomly choose a value of each of the selected parameters. A model is then built with these parameters and, using cross-validation, the quality of the model is be measured. After 10 iterations (i.e. 10 random draws), the choice will be made on the best parameters (i.e. those which provide a minimal cross-validation error) to build the final model.
- **Grid:** Activate the parameters you want to set and then select the ranges of values for each parameter. The algorithm builds a set of possible models. The best combination of parameters (the one with the lowest cross validation error) is used to build the final model.

Minimum node size:

- **Minimum parent size**: Enter the minimum number of objects that a node must contain to be split.
- **Minimum son size**: Enter the minimum number of objects that every newly created node must contain after a possible split in order to allow the splitting.

Maximum tree depth: Enter the maximum tree depth.

Cross Validation: This option allows you to quantify the quality of the model. The validation technique used to check the consistency of the classification model is the K-fold cross validation. Data is partitioned into k sub samples. Among the k subsamples, a single subsample is retained as the validation data to test the model, and the remaining $k-1$ subsamples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data. The k results can then be averaged to produce a single estimation. This option is used by default for **Automatic** and **Grid**, where $k = 5$.

- **Number of folds**: Enter the number of folds to use in the cross validation procedure.

Classes weight correction: If the number of observations for the various classes for the dependent variables are not uniform, there is a risk of penalizing classes with a low number of observations in establishing the model. To get over this problem, XLSTAT has two options:

- **Automatic**: Correction is automatic. Artificial weights are assigned to the observations in order to obtain classes with an identical sum of weights.
- **Corrective weights**: You can select the weights to be assigned to each observation.

Prior probabilities Type (QUEST only): Choose the prior probabilities type you want to include in the tree building process.

Significance level (%) (Quest Only): Enter the significance level. This value is compared to the p-values of the F and Chi-square tests. p-values smaller than this value authorize a split. This

option is not active for the CART method.

Complexity parameter: Enter the value of the complexity parameter (CP).

- **CHAID** tab:

CHAID options: These options are active when the CHAID methods are selected.

- **Merge threshold**: Enter the value of the merge significance threshold. Significance values smaller than this value lead to merge two subgroups of categories. The categories of a qualitative explanatory variable may be merged to simplify the computations and the visualization of results.
- **Authorize redivision**: Activate this option if you want to allow that previously merged categories are split again.
- **Split threshold**: Enter the value of the split significance threshold. P-values lower than this value lead to split the categories or group of categories into two subgroups of categories.
- **Bonferroni correction**: Activate this option if you want to use a Bonferroni correction during the computation of the p-value of merged categories.
- **Number of intervals**: This option is only active if quantitative explanatory variables have been selected. You can choose the maximum number of intervals generated during the discretization of the quantitative explanatory . The maximum value is 10.
- **Significance level (%)**: Enter the significance level. This value is compared to the p-values of the F and Chi-square tests. p-values smaller than this value authorize a split. This option is not active for the CART method.

Validation tab:

Validation: Activate this option if you want to use a sub-sample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.
- **Group variable**: if you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: Activate this option if you want to select data to use them in prediction mode. If this option is activated, you need to make sure that the prediction dataset is structured as the learning dataset: same variables with the same order in the selections. If variable labels option is active for the learning data, the first row of the selections listed below must correspond to the variable labels.

Quantitative: Activate this option to select the quantitative explanatory variables.

Qualitative: Activate this option to select the qualitative explanatory variables.

Observations labels: Activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2, …).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the correlation matrix.

Tree structure: Activate this option to display a table of the nodes, and information on the number of objects, the p-value of the split, the parent node and the child nodes.

Node frequencies: Activate this option to display the absolute and relative frequencies of the different nodes.

Rules: This table displays the rules in natural language, only for the dominant categories of each node.

Results by object: Activate this option to display for each observation, the observed category, the predicted category, and, in the case of a qualitative dependent variable, the probabilities corresponding to the various categories of the dependent variable.

Confusion matrix: Activate this option to display the table showing the numbers of well and badly classified observations for each of the categories.

Charts tab:

Tree chart: activate this option to display the classification and regression tree graphical. Pruning can be done by the help of the context menu of the tree chart.

- **Bar charts**: choose this option so that on the tree, the relative frequencies of the categories are displayed using a bar chart.
- **Frequencies**: activate this option to display the frequencies on the bar charts
- **%**: activate this option to display the % (of the total population) on the bar charts.
- **Pie charts**: Choose this option so that on the tree, the relative frequencies of the categories are displayed using a pie chart.

Contextual menu for the trees (only on excel 2003):

When you click on a node on a classification tree, and then do a right click on the mouse, a contextual menu is displayed with the following commands:

Show the entire tree: select this option to display the entire tree and to undo previous pruning actions.

Hide the subtree: select this option to hide all the nodes below the selected node. Hidden parts of the tree are indicated by a red rectangle of the corresponding parent node.

Show the subtree: select this option to show all the nodes below the selected node.

Set the pruning level: select this option to change the maximum tree depth.

Reset this Menu: select this option to deactivate the context menu of the tree chart and to activate the standard menu of Excel.

Roc curve : Activate this option to display the Roc curve.

Lift curve : Activate this option to display the Lift curve.

Cumulative gain curve : Activate this option to display the cumulative gain curve.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables. For qualitative variables the categories with their respective frequencies and percentages are displayed.

Correlation matrix: this table displays the correlations between the explanatory variables.

Tree structure: this table displays the nodes and information on the number of objects, the significance level of the split, and the two first child nodes. In the case of a qualitative dependent variable the predicted category is displayed. In the case of a quantitative dependent variable the expected value of the node is displayed.

Split level chart: this chart shows the significance level of the split variables for the internal nodes of the tree.

Tree chart: a legend is first displayed so that you can identify which color corresponds to which category (qualitative dependent variable). The graphical visualization of the tree allows to quickly see how it has been iteratively built, in order to obtain rules that are as pure as possible, which means that the leaves of the tree should ideally correspond to only one category.

Every node is displayed as a bar chart or a pie chart. For the pie chars, the inner circle of the pie corresponds to the relative frequencies of the categories (or intervals) to which the objects contained in the node correspond. The outer ring shows the relative frequencies of the categories of the objects contained in the parent node.

The node identifier, the number of objects, their relative frequency, and the purity (if the dependent variable is qualitative), or the predicted value (if the dependent variable is quantitative) are displayed beside each node. Between a parent and a child node, the split variable is displayed with a grey background. Arrows point from this split variable to the son nodes. The values (categories in the case of a qualitative explanatory variable) corresponding to each son node are displayed in the top left box displayed next to the son node.

Pruning can be done using the contextual menu of the tree chart. Select a node of the chart and click on the right button of the mouse to activate the context menu. The available options are described in the contextual menu section.

Node frequencies: This table displays the frequencies of the categories of the dependent variable.

Rules: The rules are displayed in natural language for the dominant categories of each node. If the option "all categories" is checked in the dialog box, then the rules for all categories and every node are displayed.

Results of search for the best model : This table displays all the results obtained when searching for parameters. The best parameters are displayed in bold.

Results by object: This table displays for each observation, the observed category, the predicted category, and, in the case of a qualitative dependent variable, the probabilities corresponding to the various categories of the dependent variable.

Confusion matrix(classification only): This table displays the numbers of well- and badly-classified observations for each of the categories (see the description section for more details).

Example

A tutorial on how to use classification and regression trees is available on the XLSTAT Help Center: http://www.xlstat.com/demo-dtr.htm

References

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.

K Nearest Neighbors

Use this tool to predict the value or the category to which belongs an observation described by a set of variables (or predictors), based on the categories or mean value of its k nearest neighbors, which are k observations for which the category or value is known and which are described by the same set of variables.

In this section:

Description

The K Nearest Neighbors (KNN) method aims to categorize observations of a prediction set whose class is unknown given their respective distances to points in a learning set (i.e. whose class is known a priori).

A simple version of **KNN** is an intuitive supervised classification approach, it can be regarded as an extension of the nearest neighbor method (NN method is a special case of KNN where $k = 1$).

The KNN classification approach assumes that each example in the learning set is a random vector in $\mathbb{R}^n$. Each point is described as $x = \{a_1(x), a_2(x), \ldots, a_n(x)\}$ where $a_r(x)$ denotes the value of the r-th attribute. $a_r(x)$ can be either a quantitative or a qualitative variable.

To determine the class of the query point x_q, each of the k nearest points $x_1, \ldots, x_k$ to x_q proceed to voting. The class of x_q corresponds to the majority class.

The following algorithm describes the basic **KNN** method:

Given a set L of size N of pre-classified samples (examples in a learning set):

$$L = \{(x_1, f(x_1)), ..(x_2, f(x_2)), ..(x_N, f(x_N)\}$$

Where $f(x_i)$ is a real value function which denotes the class of x_i

$f(x_i) \in V$ where $V = \{v_1, v_2, \ldots, v_s\}$

Given a query point or a sample to be classified x_q, let $x_1, x_2, \ldots, x_k$ be the nearest pre-classified points with a specific distance function to x_q, we have:

$$f(x_q) = \text{argmax}_{v \in V} \left(\sum_{i=1}^{K} \delta(v_i, f(x_i)) \right)$$

Where $\delta(a, b) = \begin{cases} 1 \text{ if } a = b \\ 0 \text{ if } a \neq b \end{cases}$

Origins

Nearest neighbor rules have traditionally been used in survey pattern recognition (Nilsson, 1965).

This method has also been used in several areas such as:

- Bioinformatics
- Image processing
- Computer vision
- Pattern recognition (such as handwritten character recognition)
- GIS (Geographical Information System): finding the closest cities to given positions.
- Generally, in learning systems, when the problem involves finding the nearest point or the k nearest points to a given query point.

Quantifying the similarity / dissimilarity between the query point and points in the Learning set:

The measure of dissimilarity between a given query point and the learning set is computed using a distance function. We recall that a distance function d on a set X $d : X \to R$ needs to satisfy the metric conditions:

- $d(x, y) = d(y, x)$. Symmetry property.
- $d(x, y) \geq 0$. Non-negativity property.
- $d(x, y) = 0 \Leftrightarrow x = y$. Coincidence axiom.
- $d(x, y) = d(x, z) + d(z, y)$. Triangular inequality.

Asymptotic result regarding the convergence of the basic KNN

The result established by Cover and Hart (1966) guarantees the existence of the k nearest neighbors. Let x and $x_1, \ldots, x_N$ be independent identically distributed random variables taking values in a separable metric space X. Let x'_n denote the nearest neighbor to x from the set $\{x_1, \ldots, x_N\}$.

Then $x'_n \to x$ with probability one (Cover and Hart 1966).

Complexity Of the basic KNN Method

In order to find the K nearest neighbors to a given query point, the algorithm needs to compute all the distances separating the query point to each point in the learning set. As a result, the algorithm computes N distances where N is the number of the points in the learning set. Finding the K nearest neighbors requires sorting these N distances. Consequently, the real bottleneck in the basic **KNN** algorithm resides in the sorting step. Therefore, the complexity of the basic **KNN** is in the order of $Nlog(N)$.

Quantitative metrics (distances):

Each point is considered as a quantitative vector whose components are quantitative random variables.

Many quantitative distances can be used such as:

- **Euclidean**: $d(x,y) = \sum_{i=1}^{n}(x_i - y_i)^2$
- **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**: $d(x,y) = \sum_{i=1}^{n}\frac{|x_i - y_i|}{|x_i| + |y_i|}$

Qualitative distances:

Each point is regarded as a vector whose components are all qualitative variables. When dealing with qualitative vectors, quantitative distances cannot be used. Therefore, several qualitative distances have been introduced:

The Overlap Metric OM

The Overlap Metric can be considered as basic qualitative distance:

Two vectors x and y are closer if their attributes are similar (their attributes take the same category/value). The distance between two vectors x and y can be defined as:

$$d(x, y) = \sum_{i=1}^{N} \delta(a_i(x), a_i(y))$$

Where $a_i(x)$ and $a_i(y)$ correspond to the i-th attributes of the vectors x and y.

The Value difference distance (VDM)

VDM was introduced by Graig Stanfil and David Waltz (1986). In VDM, two attributes are closer if they have the same classification class. The VDM distance between vectors x and y is given by:

$$N1 : vdm_{normalized_a}(x, y) = \sum_{c=1}^{C} |P(c|a_i(x)) - P(c|a_i(y))|^q$$

Where: - C is the total number of classes.

- $P(c|a_i(x))$: Provided $a_i(x)$, the probability that $a_i(x)$ is classified into c.
- $P(c|a_i(y))$: Provided $a_i(y)$, the probability that $a_i(y)$ is classified into c.
- q generally equals 1 or 2.

$P(c|a_i(x))$ and $P(c|a_i(y))$ are computed as follows:

$$P(c|a_i(x)) = \frac{N(a_i, x, c)}{N(a_i, x)}$$
$$P(c|a_i(y)) = \frac{N(a_i, y, c)}{N(a_i, y)}$$

Where:

- $N(a_i, x, c)$: number of instances of x for a_i in c.
- $N(a_i, x)$: number of instances of x in the data set.
- $N(a_i, y, c)$: number of instances of y for a_i in c.
- $N(a_i, y)$: number of instances of y in the data set.

Remark: Although defined for nominal attributes, the VDM distance can also be used to evaluate the distance between numeric attributes.

Computing similarity using kernels or kernel trick

Kernels can be regarded as a generalization of distance measures. They can be represented using a Hilbert space (Scholkopf 2001).

The complexity behind the computation of kernels is almost as similar as, but sometimes slightly higher than the computation involved in quantitative metrics.

Gaussian Kernel:

$$k(x, y) = \exp(-\frac{\|x - y\|}{2\delta^2})$$

Laplacian Kernel:

$$k(x, y) = \exp(-\frac{\|x - y\|}{\delta})$$

Logarithmic Kernel:

$$k(x, y) = -\log(\|x - y\|^d + 1)$$

Power kernel:

$$k(x, y) = -\|x - y\|^d$$

Sigmoid kernel:

$$k(x, y) = \tanh(\alpha x^T y + c)$$

$x^T y$ is a dot product

Linear kernel

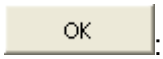
$$k(x, y) = \alpha x^T y + c$$

$x^T y$ is a dot product

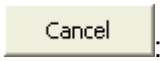
Where x, y are two vectors in $\mathbb{R}^n$; δ and d are scalars in $\mathbb{R}$.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: click this button to start the calculations.



: click this button to close the dialog box without doing any calculations.

: click this button to display help.


: click this button to reload the default options.


: click this button to delete the data selections.


: click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Training set

Y / Qualitative variables:

Qualitative: Select the response (or dependent) variable(s) you want to model. These variables must be qualitative. If several variables have been selected, XLSTAT carries out calculations for each variable separately. If a column header has been selected, make sure that the "Variable labels" option has been activated.

X / explanatory variables (training set):

Quantitative: Select the quantitative explanatory variables from learning set in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Select the qualitative explanatory variables from the learning set in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Weights: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the " Variable labels" option is activated.

Options tab:

- **General** tab :

  Model: Distance or Kernel:

  Select the way to compute the similarity between the prediction set and the learning set. The input parameter can be set to either metrics or kernels.

  Depending on the nature of the variables of the input set, either qualitative or quantitative distances can be selected:

  Quantitative distances:

  - Euclidean
  - Minkowski
  - Manhattan
  - Tchebychev
  - Canberra

```
Qualitative distances:

  * Overlap distance

  * Value difference metric


  The Kernel option enables the use of kernel functions to compute the
```

similarity between query points and points in the learning set:

```
  * Gaussian kernel

  * Laplacian kernel

  * Spherical kernel
```

* Linear kernel

  * Power kernel

In the case of the kernels option, computations are slightly longer due to the projection of points into a higher dimensional space.

  Breaking ties:

The majority voting procedure leads to the election of the query point classes. Sometimes, more than one points win the majority. This leads to a tie.

There are several ways to break ties for a given query point depending on the KNN implementation. You can break ties by selecting the options below.

  * **Random breaker**: chooses the class corresponding to a random point drawn

  * **Smallest Index**: uses the class corresponding to the first point encounte

Weighted vote: Setting the weighted vote option allows you to choose the inverse distance or the squared inverse distance as a weight for each vote of the nearest neighbors.

Observations to track: Activate this option if you want to explore which are the k nearest neighbors for all or a subset of the observations of the prediction set. * **Neighbors** tab :

Number of neighbors: Select the number of neighbors used during the KNN classification process. * **User Defined** : User can manually defines the number of neighbors, in the **Number** field.

    * **Cross Validation** : This tab allows you to quantify the quality of the
* **Automatic** : This tab allows the user to determinate the optimal number of

Prediction tab:

Prediction set:

Select the quantitative / qualitative explanatory variable data you want to use to do predictions using KNN classification. The number of variables must be equal to the number of explanatory variables in the training set.

Quantitative variables: Select the quantitative explanatory variables from learning set in the Excel worksheet. The data selected must be of type numeric. If the variable header has been

selected, check that the "Variable labels" option has been activated.

Qualitative variable: Select the qualitative explanatory variables from the learning set in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Variable labels: Activate this option if the first row of the data selections (explanatory variables, observations labels) includes a header.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

Ignore missing data : Activate this option to ignore observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest Neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected samples before and after imputation.

Results by class: Activate this option to display a table giving the statistics and the objects for each of the classes.

Results by object: Activate this option to display a table giving the class each object (observation) is assigned to in the initial object order.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

Example

A tutorial on how to use K nearest neighbors is available on XLSTAT Help Center:

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

References

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.

Naive Bayes classifier

Use this method to predict the category to which belongs an observation described by a set of quantitative and qualitative variables (predictors).

In this section:

Description

The Naive Bayes classifier is a supervised machine learning algorithm that allows you to classify a set of observations according to a set of rules determined by the algorithm itself. This classifier has first to be trained on a training dataset that shows which class is expected for a set of inputs. During the training phase, the algorithm elaborates the classification rules on this training dataset that will be used in the prediction phase to classify the observations of the prediction dataset. Naive Bayes implies that classes of the training dataset are known and should be provided hence the supervised aspect of the technique.

Historically, the Naive Bayes classifier has been used in document classification and spam filtering. As of today, it is a renowned classifier that can find applications in numerous areas. It has the advantage of requiring a limited amount of training to estimate the necessary parameters and it can be extremely fast compared to some other techniques. Finally, in spite of its strong simplifying assumption of independence between variables (see description below), the naive Bayes classifier performs quite well in many real-world situations which makes it an algorithm of choice among the supervised Machine Learning methods.

At the root of the Naive Bayes classifier is the Bayes' theorem with the "naive" assumption of independence between all pairs of variables/features. Given a class variable y and a set of independent variables $x_1, \ldots, x_n$, the Bayes' theorem states that:

$$P(y|x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n|y)}{P(x_1, \ldots, x_n)}.$$

From the naive independence assumption, the following relationship can be derived:

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y).$$

For all i, this relationship leads to:

$$P(y|x_1,\ldots,x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1,\ldots,x_n)}.$$

Since $P(x_1,\ldots,x_n)$ is constant given the input, it is regarded as a normalization constant. Thus, we can use the following classification rule:

$$P(y|x_1,\ldots,x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y),$$

and

$$\hat{y} = \max_{y}\left(P(y)\prod_{i=1}^{n} P(x_i|y)\right).$$

We can use a Maximum A Posteriori (MAP) estimation to estimate $P(y)$ and $P(x_i|y)$. Where $P(y)$ is the relative frequency of class y in the training set. Several Naive Bayes classifiers might be considered depending on the assumptions made regarding the distribution of $P(x_i|y)$.

$P(x_i|y)$ can be assumed to follow a normal distribution, in which case it has the following expression:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

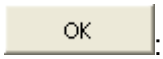
It can also be assumed to follow a Bernoulli distribution or any of the following parametric distributions available in the XLSTAT software: Log- Normal, Gamma, exponential, logistic, Poisson, binomial, uniform. In any of this cases, the distribution parameters are estimated using the moment method.

If the distribution is not known or if you are using qualitative data, XLSTAT offers the possibility to estimate an empirical distribution from the ratio of the given observation to the total number of observation for a given class y.

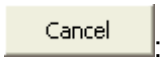
If an empirical distribution is used, it might be desirable to use a Laplace smoothing in order to avoid the null probability. This might come in handy, for instance, if a qualitative variable from the prediction data set takes a value that hasn't been met in the training phase of the algorithm. The corresponding conditional probability $P(x_i|y)$ would then be equal to 0 for every class Y_i of y, leading to the meaningless classification of the observation. In such a case, the Laplace smoothing would have the virtuous property to assign a low, but not null, conditional probability $P(x_i|y)$ to the corresponding variable. Allowing the remaining variables to be considered nonetheless to affect a class to the observation.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: click this button to start the calculations.



: click this button to close the dialog box without doing any calculations.

: click this button to display help.


: click this button to reload the default options.


: click this button to delete the data selections.


: click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Training set

Qualitative: Select the response variable(s) you want to model. These variables must be qualitative. If several variables have been selected, XLSTAT carries out calculations for each variable separately. If a column header has been selected, check that the "Variable labels" option has been activated.

X / explanatory variables (training set):

Quantitative: Select the quantitative explanatory variables from learning set in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Select the qualitative explanatory variables from the learning set in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Options tab:

Same parametric/Empirical Distribution for all quantitative variables: this option allows you to choose the same parametric/empirical distribution for all quantitative variables.

Select a specific distribution for each quantitative variable: this option allows you to select for each quantitative variable a specific parametric distribution or to consider it as an empirical distribution.

The parametric distribution can be selected from the following set of distributions: Normal, log-Normal, Gamma, exponential, logistic, Poisson, Binomial, Bernoulli, Uniform.

The qualitative variables are implicitly drawn from independent empirical distributions.

The parameters of the selected parametric distributions are estimated using the moment method.

Breaking ties:

Prediction using the naive Bayes approach can end up in a case where some classes have the same probability $P(y)$. There are several ways to break ties for a given prediction. The following options are available:

- **Random breaker**: chooses a random class in the set of classes having the same $P(y)$.
- **Smallest Index**: chooses the first class encountered in the set of classes having the same $P(y)$.

Laplace smoothing parameter:

The Laplace smoothing prevents from getting probabilities equal to zero or one.

The Laplace smoothing parameter θ is a positive real number added to the computation of the probability mass function $P(X_n = k)$ as follows:

$$P(X_n = k) = \frac{n_k + \theta}{\sum_k n_k + \theta|V|},$$

where X_n is either a qualitative or a discrete quantitative variable.

The support of X_n: V is considered to be finite; the size of V is $|V|$.

Validation tab:

The validation technique used to check the consistency of the Naive Bayes classification model is the **K-fold cross validation technique**. Data is partitioned into k subsamples of equal size. Among the k subsamples, a single subsample is retained as the validation data to test the model, and the remaining $k-1$ subsamples are used as training data. k can be specified in the **number of folds** field.

Prediction tab:

Prediction set: Activate this option if you want to select data to use them in prediction mode. If this option is activated, you need to make sure that the prediction dataset is structured as the learning dataset: same variables with the same order in the selections. If variable labels option is active for the learning data, the first row of the selections listed below must correspond to the variable labels.

Quantitative variable: Select the quantitative explanatory variables from learning set in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative variable: Select the qualitative explanatory variables from the learning set in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Variable labels: Activate this option if the first row of the data selections (explanatory variables, observations labels) includes a header.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest Neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Results by class: Activate this option to display a table giving the statistics and the objects for each of the classes.

Results by object: Activate this option to display a table giving the class each object (observation) is assigned to in the initial object order.

Posterior probabilities of each class: Activate this option to display the table which summarizes the posterior probabilities corresponding to each class $P(Y = y)$ for all predicted observations.

Confusion matrix: Activate this option to display the confusion matrix. The confusion matrix contains information about observed and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. Diagonal values correspond to correct predictions. The higher the sum of the diagonal values according to the total the better the classifier.

Accuracy of the model: Activate this option to display model accuracy, which is the proportion of correct predictions.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

Results corresponding to the descriptive statistics of the learning set:

The number of observations corresponding to each variable in the learning set, its mean (in case of quantitative variable or modes in case of qualitative variable) and its standard deviation.

Results corresponding to the parameters involved in the classification process:

The kind of probability distribution is reported.

The qualitative variables are considered to follow implicitly an empirical distribution.

The nature of the a priori distribution of the classes (uniform, not uniform) is also reported.

Results regarding the classifier

In order to evaluate and to score the naive Bayes classifier, a simple confusion matrix computed using the leave one out method as well as an accuracy index are displayed.

Results regarding the validation method

The error rate of the naive Bayes model obtained using the K folded-cross validation is reported. The number of folds is also reported to the user.

The cross validation results enables the selection of the adequate model parameters.

Result corresponding to the predicted classes

The predicted classes obtained using the naive Bayes classifier are displayed. In addition to the predicted classes, the a posteriori probabilities used to predict each observation are also

reported.

Example

A tutorial on how to use the naive Bayes classifier is available on the XLSTAT Help Center:

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

References

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.

Support Vector Machine

Use this method to perform a binary classification, a multiclass classification or a regression on a set of observations described by qualitative and/or quantitative variables (predictors).

In this section:

Description

The Support Vector Machine (SVM) was invented in the context of the statistical learning theory (Vapnik and Chervonenkis, 1964). It was not until the mid-90s that an algorithm implementation of the SVM was proposed with the introduction of the kernel trick (Boser, B., Guyon, I., & Vapnik, V., 1992) and the generalization to the non separable case (Cortes, C. & Vapnik V. 1995). Since then, the SVM has known numerous developments and gained popularity in various areas such as Machine Learning, optimization, neural networks or functional analysis. It is one of the most successful learning algorithms. Its ability to compute a complex model at the price of a simple one made it a key component to the Machine Learning domain where it has become famous in applications such as text or image recognition.

Binary classification

The SVM aims to find a separation between two classes of objects with the idea that the larger the separation, the more reliable the classification. In its simplest form, the linear and separable case, the algorithm will select a hyperplane that separates the set of observations into two distinct classes in a way that maximizes the distance between the hyperplane and the closest observation of the training set.

Suppose the optimal separating hyperplane P_0 is known. P_0 is given by the following equation:

$$P_0 : x^T \cdot w - b = 0$$

Where x^T is the set of predictors of the observation, W the normal vector to the hyperplane, b the origin of the hyperplane.

As the training dataset is assumed to be separable, we can identify two hyperplanes, namely P+ and P-, parallel the separation hyperplane such that:

$$P_+ : x^T \cdot w - b = 1$$
$$P_- : x^T \cdot w - b = -1$$

Where $y_i = \pm 1$ denotes the two possible outcomes of the output class. The distance between P_+ and P_- , $\frac{2}{\|w\|}$, is called the margin. This is the parameter we wish to maximize during our optimization in order to ensure the largest possible margin.

We can then rephrase our optimization problem as:

$$min_{w,b} \|w\|$$

subject to:

$$y_i(x_i^T \cdot w - b) \geq 1, i = 1, ..., N$$

The w and b resulting from this minimization defines our classifier. It is worth noting that only points lying near the boundary define the hyperplane. Those observations are called support vectors and are of special interest as they define the classifier. On the contrary, observations well inside their class boundary only have a marginal impact, this might be a useful property in situations where some outliers are present in the training dataset.

In the case where classes overlap, data are no longer separable and some points must be allowed on the wrong side of the margin. A slack variable ϵ_i must be introduced to account for the amount by which the prediction falls on the wrong side. Our optimization problem becomes:

$$min_{w,b} \|w\|$$

subject to:

$$\begin{cases} y_i \left(x_i^T . w - b \right) \geq 1 - \epsilon_i, \forall i \\ \epsilon_i \geq 0, \sum_{i=1}^{N} \epsilon_i \leq K \end{cases}$$

Misclassification occurs when $\epsilon_i > 1$ and the total amount of misclassification, $\sum \epsilon_i$, is bounded by the constant K. In terms of computation, it is more convenient to rephrase the above expression as the equivalent form that follows:

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \epsilon_i$$

subject to:

$$\begin{cases} y_1(x_i^T \cdot w - b) \geq 1 - \epsilon_i, \forall i \\ \epsilon_i \geq 0 \end{cases}$$

Where the regularization parameter C is introduced as a replacement of K. Intuitively, this regularization parameter reflects how important misclassification is. A large C means that we want to limit misclassification at the price of a narrower margin. The extreme case being the separable case where $C = \infty$. A smaller C means that relatively more misclassifications are authorized with the benefit of a larger margin.

On a computational aspect, the above optimization problem has a quadratic form with linear inequality constraints. It can then be solved using the method of Lagrange multipliers.

The Lagrange primal function is:

$$L_p = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\epsilon_i - \sum_{i=1}^{N}\alpha_i[y_i(x_i^T w + b) - (1-\epsilon_i)] - \sum_{i=1}^{N}\mu_i\epsilon_i$$

Which is minimized with respect to w_i, b and ϵ_i.

Setting the respective derivatives to zero gives:

$$\begin{aligned} w &= \sum_{i=1}^{N}\alpha_i y_i x_i \\ 0 &= \alpha_i y_i \\ \alpha_i &= C - \mu_i \end{aligned}$$

And the positive constraints $\alpha_i, \mu_i, \epsilon_i \geq 0, \forall i$.

The wolf dual objective function to be maximized is:

$$L_D = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\alpha_i\alpha_{i'}y_i y_{i'}x_i^T x_{i'}$$

subject to:

$$\begin{cases} 0 \leq \alpha_i \leq C \\ \sum_{i=1}^{N}\alpha_i y_i = 0 \end{cases}$$

The Karush-Kuhn-Tucker (KKT) conditions include the following additional constraints:

$$\begin{cases} \alpha_i[y_i(x^T w + b) - (1-\epsilon_i)] = 0 \\ \mu_i\epsilon_i = 0 \\ y_i(x^T w + b) - (1-\epsilon_i) \geq 0 \end{cases}$$

for $i = 1, ..., N$.

In this non-separable case, support vectors are identified as the observations for which $\alpha_i > 0$. Those lying exactly on the border have $0 < \alpha_i < c$ while for the reminders $\alpha_i = C$.

The SVM classifier can now be extended to non-linear classification by using kernel functions. This method known as the kernel trick is similar to the one used in logistic regression where the linear method is made more flexible by enlarging the input feature space. In SVM, the usage of

kernels allows the feature space to get very large. More complicated structures can then be detected. XLSTAT proposes 3 kernels in addition to the linear approach:

- Power kernel: $k(x_i, x_i') = (\gamma \cdot (x_i^T x_i') + coefficient)^{degree}$
- Radial Basis Function (RBF) kernel: $k(x_i, x_i') = e^{-\gamma \|x_i - x_i'\|^2}$
- Sigmoid kernel: $k(x_i, x_{i'}) = \tanh(\gamma \cdot (x_i^T x_i' + coefficient))$

Once the basis function is chosen, the procedure remains identical to the one described above.

The optimization problem discussed above is solved using the Sequential Minimal Optimization (SMO) as proposed by John Platt (Platt J., 1998). This algorithm breaks the problem into smaller subproblems that can be solved analytically. The computation burden is dramatically reduced and the SVM classifier becomes an extremely powerful classifier for a very limited computational cost. Nevertheless, a version of the SVM to achieve fast convergence was proposed by Fan *et al.* using Second Order information (Fan, R., Chen, P. & Lin, C., 2005).

Multiclass classification

SVM can only resolve binary problems then different methods have been developed to solve multiclass problems. They all use the same principle: transform the multiclass problem into several binary problems. XLSTAT proposes two different methods to solve multiclass problems.

The first available method in XLSTAT is called One versus all. Let K denote the number of classes and ω_i denote the ith class (with $i \in \{1, \ldots, K\}$). The ith classifier output function f_i is trained taking the examples from ω_i as positive and the examples from all other classes as negative. For a new example x, the class with the largest value of $f_i(x)$ is assigned.

The second available method in XLSTAT is called "One versus one with max-wins strategy". A binary classifier is created for every pair of distinct classes. Consequently, $K(K-1)/2$ binary classifiers are constructed. The binary classifier C_{ij} is trained taking the examples from ω_i as positive and the examples from ω_j as negative. For a new example x, if classifier C_{ij} classifies x in class ω_i, then the vote for class ω_i is added by one. Otherwise, the vote for class ω_j is increased by one. After each of the $K(K-1)/2$ binary classifiers vote, the class with the largest number of votes is assigned to x. If a tie occurs, the class of x is designated randomly.

Regression

SVM method was generalized to be applied to regression problem or time series prediction. Let the training set $\{x_i, y_i\}$ for $i = 1, ..., N$ where x is the set of predictors of the observation and $y_i \in \mathbb{R}$.

In the linear case, the goal is to estimate f, with a deviation less than ϵ compare to target variable y. f is represented by:

$$ f(x) = x ^ { T } \cdot w + b $$ Where w is the normal vector to the hyperplane and b the origin of the hyperplane.

The optimization problem can be formuled this way:

$$min_{w,b} \frac{1}{2}\|w\|^2$$

subject to:

$$\begin{cases} y_i - w^t \cdot x_i - b \leq \epsilon \\ w^t \cdot x_i + b - y_i \leq \epsilon \end{cases}$$

We introduced slack variables ξ and ξ^* to cope with infeasible constraints, and we arrived to this formulation:

$$min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$

subject to:

$$\begin{cases} y_i - w^t \cdot x_i - b \leq \epsilon + \xi_i \\ w^t \cdot x_i + b - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

Where $C > 0$ is the parameter of regularization, the bigger C is, the less deviations are penalized. If $C = 0$, there is no penalty.

After we add slack variables, we have to deal with a so called ϵ-insensitive loss function $|\xi|_\epsilon$ described by:

$$|\xi|_\epsilon = \begin{cases} 0 & if\ |\xi| \leq \epsilon \\ |\xi| - \epsilon & otherwise \end{cases}$$

Again, as for the classification method, we solve the above optimization problem using the method of Lagrange multipliers.

The Lagrange primal function is:

$$\begin{aligned} L_p = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) - \sum_{i=1}^{N}\alpha_i[\epsilon + \xi_i - y_i + w^t.x_i + b] \\ - \sum_{i=1}^{N}\alpha_i^*[\epsilon + \xi_i^* + y_i + w^t.x_i - b] - \sum_{i=1}^{N}(\eta_i\xi_i + \eta_i^*\xi_i^*) \end{aligned}$$

Where $\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geq 0,\ \forall i$ are the Lagrange multipliers

We minimize L with respect to b, w_i, ξ_i and ξ_i^* and setting the respective derivatives to zero gives:

$$\sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0$$

$$w - \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)x_i = 0$$

$$C - \alpha_i - \eta_i = 0$$

$$C - \alpha_i^* - \eta_i^* = 0$$

We can eliminated the variables η and η^* through:

$$\eta = C - \alpha_i$$

$$\eta^* = C - \alpha_i^*$$

Finally, the dual objective function to be maximized is:

$$ \begin{aligned} L *{ D } &= - \frac { 1} { 2} \sum { i,j = 1} ^ { N }(\alpha_{i} - \alpha_{i}^{(*})(\alpha_{j}-\alpha_{j})(x_{i}^{t} \cdot x_{j})) \

&-\epsilon\sum *{ i ^ = 1} ^ { N } (\alpha* { i }+ \alpha *{ i} ^ { *}) + \sum* { i ^ = 1} ^ { N }y *{ i }(\alpha* { i }+ \alpha _ { i} ^ { *}) \end{aligned} $$

subject to:

$$\begin{cases} \sum_{i=1}^{N}(\alpha_i + \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \end{cases}$$

The Karush-Kuhn-Tucker (KKT) conditions include the following additional constraints:

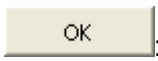
$$ \left\lbrace \begin{aligned} & \alpha *{ i } [\epsilon* + \xi+ w^{T} \cdot x_{i}+b] = 0 \ & \alpha_ { i }^{} [\epsilon* + \xi_{i}^{} + w^{t} \cdot x_{i}+b] = 0 \ & (C - \alpha_{i}) \xi _ { i } = 0 \ & (C - \alpha_{i}^{}) \xi _ { i }^{} = 0 \end{aligned} \right. $$ for $i = 1, ..., N$.

As for the classification method, the SVM method for regression can be extended for the non-linear cases using kernels.

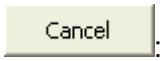
The optimization problem is also solved using the Sequential Minimal Optimization (SMO) using second order information as proposed by Fan and Al. (Fan, R., Chen, P. & Lin, C., 2005).

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Response variable: Select the response variable you want to model. If a column header has been selected, make sure the "Variable labels" option has been activated.

Response type: Select the type of response you have:

- **Qualitative**: Select this type, if you want to fit a classification model. Then, choose the type of classification: Choose **binary**, if you selected a variable containing exactly two distinct values. The positive categories or classes correspond to the first category met in the dataset and the negative classes to the second.
  If you have more than two classes, choose between **one versus one** or **one versus all** methods corresponding to a multiclass SVM (see Description section).

- **Quantitative**: If your response type contains real values, choose this type to fit a regression model.

Explanatory variables:

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If the variable header has been selected, make sure the "Variable labels" option has been activated.

Qualitative: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, make sure the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observation labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Observation weights: Activate this option if you want to use different weights for an observation, especially when class sizes are asymmetrics. If the variable header has been selected, make sure the "Variable labels" option has been activated.

Options tab:

SMO parameters: This option allows you to tune the optimization algorithm to your specific needs. There are 3 tunable parameters:

- **C**: This is the regularization parameter (see the description for more details).
- **Tolerance**: This value defines the tolerance when comparing 2 values during the optimization. This parameter can be used to speed up computations.
- **Epsilon**: Used for regression only, this parameter define the insensitive tube with a rayon equal to ϵ.

NB: The ϵ parameter for classification, is a machine dependent accuracy parameter and it is initialized to 10^{-12}.

Preprocessing: This option allows you to select the way the explanatory data are rescaled. There are 3 options available:

- **Rescaling**: Quantitative explanatory variables are rescaled between 0 and 1 using the observed minimum and maximum for each variable.
- **Standardisation**: Both qualitative and quantitative explanatory variables are standardized using the sample mean and variance for each variable.
- **None**: No transformation is applied.

Cross-validation: This option allows you to run a k-fold cross-validation to quantify the quality of the classifier or the regression with chosen parameters. Data is partitioned into k equally subsamples of equal size. A single subsample is retained as the validation data to test the model, and the remaining k-1 subsamples are used as training data.

Kernel: This option allows you to select the kernel you wish to apply to your dataset to extend the feature space. There are 4 kernels available:

- **Linear kernel**: This is the basic linear dot product.
- **Power kernel**: This kernel is detailed in the description. If you select this kernel, you have to enter the coefficient and gamma parameters.
- **RBF kernel**: This the Radial Basis Function as detailed in the description. If you select this kernel, you have to enter the gamma parameter.
- **Sigmoid kernel**: This kernel is detailed in the description. If you select this kernel, you have to enter the coefficient and gamma parameters.

Validation tab:

Validation: Activate this option if you want to use a subsample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: Activate this option if you want to select data to use them in prediction mode. If you activate this option, you need to make sure that the prediction dataset is structured like the estimation dataset: the same variables with the same order in the selections.

Quantitative: Activate this option to select the quantitative explanatory variables. The first row should include variable labels if the Variable labels option is activated on this page.

Qualitative: Activate this option to select the qualitative explanatory variables. The first row should include variable labels if the Variable labels option is activated on this page

Observation labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Variable labels" option is activated, you need to include a header in the selection. If this option is not activated, the observation labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Variable labels: Activate this option if the first row of the data selections (explanatory variables and observations labels) includes a header.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected for each of the activated dataset (training, validation, prediction).

Estimation summary: Activate this option to display a summary of the optimized SVM classifier or regressor.

List of support vectors: Activate this option to display the complete list of support vectors and their associated coefficient alpha for the classification or alpha-alpha* for the regression, as presented in the description.

Performance metrics: Activate this option to display performance indicators for the classification or the regression of the training dataset and the validation dataset (if activated).

Results by object: Activate this option to display the classification or the regression results for each observation of the training dataset, the validation dataset and the prediction dataset (if activated).

Confusion matrix: Activate this option, only in the case of classification, to display the confusion matrix for the classification of the training dataset and the validation dataset. The confusion matrix contains information about the observed and predicted classifications by the algorithm. Performances can be evaluated using the confusion matrix. The diagonal contains correct predictions. The greater the sum of elements of the diagonal, the better the classifier.

Results

Descriptive statistics:

The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

Results regarding the estimation:

A summary description of the optimized classifier or regressor is displayed. In the case of classification, the positive and negative classes are indicated. For both, the training sample size and both optimized parameters the bias and the number of support vectors are displayed.

Results regarding the list of support vectors:

A table containing the value of the class, the optimized value of alpha or alpha-alpha* in the case of regression, and the rescaled explanatory variables as they were used during the optimization is displayed for each identified support vector.

Results regarding the confusion matrices:

The confusion matrix is deduced from prior and posterior classifications together with the overall percentage of well-classified observations.

Results regarding the performance metrics:

There are 10 classification metrics displayed if this option is active:

Accuracy, Precision, Recall, F-score, Specificity, False Positive Rate (FPR), Prevalence, Cohen's kappa, Null Error Rate (NER) and Area Under Curve (AUC).

In addition to these indicators, the ROC curve is displayed for the training sample and the validation sample (if activated). It represents the evolution of the proportion of the sensitivity as a function of $1-$ the specificity.

There are 3 performance metrics displayed in the case of regression:

Mean Squared Error, Mean Absolute Error and the coefficient of determination R^2.

Indicators in the first column correspond to the training sample and those in the second column to the validation sample (if activated).

Results corresponding to the predicted classes or values:

The predicted classes or values obtained using the SVM classifier or the SVM regressor are displayed for the training, validation and prediction dataset (if activated). Moreover, in the case of binary classification, the decision function is displayed.

Results corresponding to the cross-validation:

3 performance metrics are displayed if cross-validation is active. For each k fold, classification error rate, F-score and Balanced ACcuracy are displayed in the case of binary classification.

In the case of regression, Mean Squared Error, Mean Absolute Error and coefficient of determination are displayed.

Example

A tutorial on how to use the Support Vector Machine is available on the XLSTAT Help Center:

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

https://help.xlstat.com/6471-training-support-vector-machine-regression-svr-excel

References

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.

One-class Support Vector Machine

Use this method to perform novelty detection on a set of observations described by qualitative and/or quantitative variables (predictors).

In this section:

Description

The Support Vector Machine (SVM) was invented in the context of the statistical learning theory (Vapnik and Chervonenkis, 1964). It was not until the mid-90s that an algorithm implementation of the SVM was proposed with the introduction of the kernel trick (Boser, B., Guyon, I., & Vapnik, V., 1992) and the generalization to the non separable case (Cortes, C. & Vapnik V. 1995). Since then, the SVM has known numerous developments and gained popularity in various areas such as Machine Learning, optimization, neural networks or functional analysis. It is one of the most successful learning algorithm. Its ability to compute a complex model at the price of a simple one made it a key component to the Machine Learning domain where it has become famous in applications such as text or image recognition.

One-class Support Vector Machine

It was in 1999 that Schölkopf *et al.* proposed an expansion to SVM for the unsupervised learning and more precisely for novelty detection. In the case of novelty detection, the algorithm learns on a dataset assuming that all observations are normal. Then, the constructed model will identify a new observation as an outlier or not.

The One-class Support Vector Machine (One-class SVM) algorithm seeks to envelop underlying inliers. The aim is to seperate data into two classes, the positive one considered as the class of inliers and the negative one considered as the class of outliers. Besides, most of the training data must belong to the positive class while the volume of envelope is minimal.

We want to separate data by a hyperplane whose distance from the origin $\frac{\rho}{\|w\|}$ is maximum. So, we need to solve the following quadratic problem:

$$\min_{w,\xi,\rho} \frac{1}{2}\|w\|^2 + \frac{1}{\nu l}\sum_{i=1}^{N}\xi_i - \rho$$

subject to:

$$\begin{cases} (w \cdot \Phi(x_i)) \geq \rho - \xi_i, \forall i \\ \xi_i \geq 0 \end{cases}$$

Where $\Phi : X \rightarrow F$ is a feature map, X space of data and $\nu \in [0, 1]$ is the trade-off between the number of observations in the positive class and a little $||w||^2$.

Then the decision function can be formuled as:

$$f(x) = sgn((w \cdot \Phi(x)) - \rho)$$

On a computational aspect, the above optimization problem has a quadratic form with linear inequality constraints. It can then be solved using the method of Lagrange multipliers.

The Lagrange primal function is:

$$L_p = \frac{1}{2}\|w\|^2 + \frac{1}{\nu l}\sum_{i=1}^{N}\xi_i - \rho - \sum_{i=1}^{N}\alpha_i[(w \cdot \phi(x_i)) - \rho + \xi_i] - \sum_{i=1}^{N}\mu_i\xi_i$$

Which is minimized with respect to w, ξ, ρ, α and μ.

Setting the respective derivatives to zero gives:

$$w = \sum_{i=1}^{N}\alpha_i\Phi(x_i)$$

$$\alpha_i = \frac{1}{\nu l} - \beta_i \leq \frac{1}{\nu l}, \sum_{i=1}^{N}\alpha_i = 1$$

The wolf dual objective function to be maximized is:

$$L_D = \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\alpha_i\alpha_{i'}k(x_i, x_{i'})$$

subject to:

$$\begin{cases} 0 \leq \alpha_i \leq \dfrac{1}{\nu N} \\ \displaystyle\sum_{i=1}^{N}\alpha_i = 1 \end{cases}$$

Where $k(.,.)$ is the kernel function assumed to be positive definite and defined by: $k(x, y) = (\Phi(x) \cdot \Phi(y))$.

XLSTAT proposes 3 kernels in addition to the linear approach:

- Power kernel: $k(x_i, x_i') = (\gamma \cdot (x_i^T x_i') + coefficient)^{degree}$

- Radial Basis Function (RBF) kernel: $k(x_i, x_i') = e^{-\gamma \|x_i - x_i'\|^2}$
- Sigmoid kernel: $k(x_i, x_{i'}) = \tanh(\gamma \cdot (x_i^T x_i' + coefficient))$

Finally, the decision function with a kernel expansion is:

$$f(x) = sgn(\sum_{i=1}^{N} \alpha_i k(x_i, x) - \rho)$$

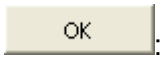
Support vectors are identified through observations with $\alpha_i > 0$.

In the case, the decision function is positive, the observation will be predicted as an inlier, in the contrary case it will be predicted as an outlier.

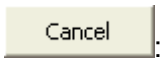
As others SVM methods available in XLSTAT the optimization problem is solved thanks to the Sequential Minimal Optimization (SMO) using second order information as proposed by Fan and Al. (Fan, R., Chen, P. & Lin, C., 2005).

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Explanatory variables:

- **Quantitative**: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If the variable header has been selected, check that the "Variable labels" option has been activated.

- **Qualitative**: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Known classes: Activate this option if you already know the classes of each observation from the training dataset and want to include it. Selected data must be binary (1, -1, -1, 1,...), then you need to choose through "Outlier class" which one of the two classes is the outlier class. If a column header has been selected, check that the "Variable labels" option has been activated.

Outlier class : This value define the outlier/negative class.

Inlier class : In the case you don't have true classes of the training dataset, you need to define the value of the inlier/positive class.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Observation weights: Activate this option if you want to use different weight for a class, specially when class sizes are asymmetrics. If the variable header has been selected, check that the "Variable labels" option has been activated.

Options tab:

SMO parameters: This option allows you to tune the optimization algorithm to your specific needs. There are 2 tunable parameters:

- **Nu**: This value is the regularization parameter and is between 0 and 1 (see the description for more details).

- **Tolerance**: This value define the tolerance when comparing 2 values during the optimization. This parameter can be used to speed up computations.

Preprocessing: This option allows you to select the way the explanatory data are rescaled. There are 3 options available:

- **Rescaling**: Quantitative explanatory variables are rescaled between 0 and 1 using the observed minimum and maximum for each variable.
- **Standardisation**: Both qualitative and quantitative explanatory variables are standardized using the sample mean and variance for each variable.
- **None**: No transformation is applied.

Cross-validation: Available only when "Known classes" is activated. This option allows you to run a k-fold cross-validation to quantify the quality of the classifier. Data is partitioned into k equally subsamples of equal size. A single subsample is retained as the validation data to test the model, and the remaining k-1 subsamples are used as training data.

Kernel: This option allows you to select kernel you wish to apply to your dataset to extend the feature space. There are 4 kernels available:

- **Linear kernel**: This is the basic linear dot product.
- **Power kernel**: This kernel is detailed in the description. If you select this kernel, you have to enter the coefficient, the degree and gamma parameters.
- **RBF kernel**: This the Radial Basis Function as detailed in the description. If you select this kernel, you have to enter the gamma parameter.
- **Sigmoid kernel**: This kernel is detailed in the description. If you select this kernel, you have to enter the coefficient and gamma parameters.

Validation tab:

Validation: Available only when "Known classes" is activated. Activate this option if you want to use a subsample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: Activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections.

Quantitative: Activate this option to select the quantitative explanatory variables. The first row should include variable labels if the Variable labels option is activated on this page.

Qualitative: Activate this option to select the qualitative explanatory variables. The first row should include variable labels if the Variable labels option is activated on this page

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Variable labels: Activate this option if the first row of the data selections (explanatory variables and observations labels) includes a header.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected for each of the activated dataset (training, validation, prediction).

Estimation summary: Activate this option to display a summary of the optimized SVM classifier.

List of support vectors: Activate this option to display the complete list of support vectors and their associated coefficient α, as presented in the description.

Results by object: Activate this option to display predictions for each observation of the training dataset, the validation dataset and the prediction dataset (if activated).

Performance metrics: Available only when "Known classes" is activated. Activate this option to display the ROC curve and performance indicators for the classification of the training dataset and the validation dataset (if activated).

Confusion matrix: Available only when "Known classes" is activated. Activate this option, only in the case of classification, to display the confusion matrix for the classification of the training dataset and the validation dataset. The confusion matrix contains informations about the observed and predict classifications by the algorithm. Performances can be evaluated by the use of this confusion matrix. The diagonal contains correct predictions. The greater the sum of elements of the diagonal are, the better the classifier is.

Results

Descriptive statistics:

The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

Results regarding the estimation:

A summary description of the optimized classifier is displayed. The outlier class, the training sample size are displayed. Also, both optimized parameters bias corresponding to ρ and the number of support vectors are displayed.

Results regarding the list of support vectors:

A table containing the optimized value of α, and the rescaled explanatory variables as they were used during the optimization is displayed for each identified support vector.

Results regarding the confusion matrices:

The confusion matrix is deduced from prior and posterior classifications together with the overall percentage of well-classified observations.

Results regarding the performance metrics:

There are 10 classification metrics displayed if this option is active:

Accuracy, Precision, Recall, F-score, Specificity, False Positive Rate (FPR), Prevalence, Cohen's kappa, Null Error Rate (NER) and Area Under Curve (AUC).

Indicators in the first column correspond to the training sample and those in the second column to the validation sample (if activated).

In addition to these indicators, the ROC curve is displayed for the training sample and the validation sample (if activated). It represents the evolution of the proportion of the sensitivity as a function of $1-$ the specificity.

Results corresponding to the predicted classes:

The predicted classes obtained using the SVM classifier are displayed for the training, validation and prediction dataset (if activated). Moreover, the decision function is displayed.

Results corresponding to the cross-validation:

3 performances metrics are displayed if cross-validation is active. For each k fold, classification error rate, F-score and Balanced ACcuracy are displayed in the case of binary classification.

Example

A tutorial on how to use the Support Vector Machine is available on the XLSTAT Help Center:

http://www.xlstat.com/demo-SVM1class.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)

Use this method to perform anomaly detection and clustering on a set of observations described by quantitative and/or qualitative variables.

In this section:

Description

DBSCAN stands for *Density-based Spatial Clustering of Applications with Noise* proposed by Ester, Kriegel, Sander and Xu in 1996. It is the most widely used unsupervised learning method among density-based clustering methods. There are several advantages to using this type of method: the ability to create an unknown number of classes, to create classes with non-convex shapes and the ability to handle anomalies.

In order to use the DBSCAN method, 2 parameters are required:

- $\epsilon > 0$
- The minimum number of points, also called $MinPts > 0$

Several definitions allow us to understand how classes are made. First, we must define and count neighbors for each point. A neighbor is defined as any point p from the training dataset with a distance less than or equal to ϵ from a point q.

Note that by definition the point q is its own neighbor.

3 types of points can be defined with the DBSCAN algorithm:

- Core point: A point with as many or more neighbors as the minimum number of points.
- Border point: A point that has fewer neighbors than the minimum number of points but is a neighbor of a core point.
- Noise point: Neither a core point nor a border point.

A point p is **directly density-reachable** from q, if q is a core point and p a neighbor of q. A point p is **density-reachable** from q if there is an ordered sequence of points directly density-

reachable from the previous point. Two points p and q are **density-connected** if there is a point o such that both p and q are density-reachable from o.

Finally, Ester *et al.* defined a **class** as a subset of the dataset that fits two conditions:

- If p belongs to the class C and q is density-reachable from p then q belongs to C.
- All points in the class C are mutually density-connected.

The DBSCAN algorithm

The DBSCAN algorithm visits all points of the training dataset and marks them *visited* as it goes.

If a point is a core point, then the first class is started (named class 1). The core point and its neighbors are assigned to class 1. Then, the algorithm visits its neighbors in order to find another core point and assigns it to class 1. This step allows the class to expand. The algorithm stops to expand class 1 when all density-reachable points have been visited.

The algorithm continues to visit the unvisited points and will start a new class if another core point is found. This class can also be expanded and so on...

Finally, all points which are not assigned to a class are noise points.

Prediction with DBSCAN

DBSCAN allows you to predict the class of new observations.

First, it must find the neighbors of each new observation in the training dataset. If a new observation is a neighbor of a core point (from the training dataset) then the new observation is assigned to the same class as the core point.

If the new observation has no core point in its neighbors, then it is considered as a noise point

Note that the visiting order can change the class assigned to the border points during learning and prediction.

K-dimensional tree

Use the K-dimensional tree when the dataset contains only quantitative variables (Bentley, 1975). It makes it possible to not compute all the distances to find all the neighbors in a radius of size epsilon.

The k-dimensional tree is a binary tree constructed by sorting the points from one dimension and dividing the space into 2 from the median. Points with a value less than or equal to the median in this dimension are stored in the left child node while points with a value greater than the median are stored in the right child node. The construction of the tree stops when there is only one point left in a node.

Distance metrics

There are different distance metrics to compute distances with any type of variable:

There are 5 metrics when only quantitative variables are selected:

- Euclidean distance
- Minkowski distance
- Manhattan distance
- Chebychev distance
- Canberra distance.

When only qualitative variables describe the observations, the Overlap distance is used.

With mixed data, the HEOM (Heterogeneous Euclidean Overlap Metric) is used.

Measuring DBSCAN's Performance

The silhouette score measures the quality of the classification of an observation in a class. It is formulated as:

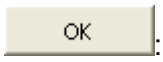
$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))},$$

where $a(i)$ is the mean distance between i and all other points in the same class, and $b(i)$ is the mean distance between i and all other points in the nearest class.

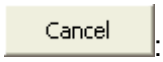
The silhouette score varies between -1 and 1 and the closer its value is to 1 the better an observation lies within its class.

Dialog box

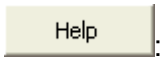
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. Below are descriptions of the various elements of the dialog box.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.



: Click this button to display help options.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.


: Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Variables:

- **Quantitative**: Activate this option if you want to include one or more quantitative variables in the model, then select the corresponding variables in the Excel worksheet. The data selected must be of the numerical type. If the variable header has been selected, make sure the "Variable labels" option has been activated.
- **Qualitative**: Activate this option if you want to include one or more qualitative variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, make sure the "Variable labels" option has been activated.

Observation weights: Activate this option if you want to use different weights for the samples. If the variable header has been selected, make sure the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet in the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observation labels) includes a header.

Observation labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Variable labels" option is activated, you need to include a header in the selection. If this option is not activated, the observation labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Options tab:

Parameters: This option allows you to tune the algorithm to your specific needs. There are 2 tunable parameters:

- **Epsilon**: This value is the maximum distance between two points for them to be considered as neighbors. In other words, if a point has a distance less than or equal to ϵ from another point, then it is in the neighborhood of the first point.
- **Minimum number of points**: This parameter defines the minimum number of points in the neighborhood of a point for it to be considered a core point (see description). An

interval may be entered in order to run several analyses with a different minimum number of points and a step size of 1.

Preprocessing: This option allows you to select the way the data are rescaled. There are 2 options available:

- **Standardization**: quantitative variables are standardized using the sample mean and variance for each variable.
- **None**: No transformation is applied.

Search method: this option allows you to choose between 2 methods in order to find the neighbors of a point:

- **K-dimensional tree**: choose this option, if only quantitative variables are entered (see description).
- **Distance matrix**: choose this option with any type of variables to use a neighbor search with the distance matrix. All distances between each point will be computed, enabling the computation of the silhouette score.

Distance: This option allows you to select the metric you wish to apply to your dataset according to the type of variables:

- For quantitative variables:
- **Euclidean distance**
- **Minkowski distance**
- **Manhattan distance**
- **Chebychev distance**
- **Canberra distance**
- For qualitative variables:
    - **Overlap distance**
- For quantitative and qualitative variables:
    - **HEOM distance**

Prediction tab:

Prediction: Activate this option if you want to select data to use in prediction mode. If you activate this option, you need to make sure that the prediction dataset is structured like the estimation dataset: the same variables with the same order in the selections.

Quantitative: Activate this option to select the quantitative variables. The first row should include variable labels if the Variable Labels option is activated on this page.

Qualitative: Activate this option to select the qualitative variables. The first row should include variable labels if the Variable Labels option is activated on this page

Observation labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Variable Labels" option is activated, you need to include a header in the selection. If this option is not activated, the observation labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Variable labels: Activate this option if the first row of the data selections (variables and observation labels) includes a header.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected for each activated dataset (training, prediction).

Correlation matrix: Activate this option to display a view of the correlations between the various variables selected.

Number of objects by class: Activate this option to display the number of observations considered as noise and/or the number of observations assigned to each class.

Results by class: Activate this option to display a table sorting the observations by class.

Results by object: Activate this option to display the class assigned to each observation from the learning sample. If the prediction option is activated, classes assigned to the new observations are always displayed.

- **Silhouette score**: Activate this option, available only if you use distance matrix as your search method, to display silhouette scores for each observation from the training sample and the associated graph.

Distance matrices: Activate this option, available only if you use distance matrix as your search method, to display the distance matrices.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables. For

qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

Correlation matrix: This table is displayed to give you a view of the correlations between the various variables selected.

Number of objects by class: This table is displayed to give you a view of the size of each class and the number of noise points.

Results regarding distance matrices: One or two distance matrices are displayed if the prediction option is activated. The first matrix shows distances between each point of the training sample. The second matrix shows distances between the new observations and the observations of the training sample.

Results regarding objects: Classes assigned to each observation using the DBSCAN algorithm are displayed for the training and the prediction sample. If the class is 0 it means the observation is considered as a noise point. In addition, the silhouette score of each observation is displayed in the second column (if the option is activated).

A graph of the silhouette scores is displayed if the option is activated. Observations are grouped by classes in descending order with respect to the silhouette coefficient.

Results associated with objects sorted by class: this table is displayed to show the observations sorted by class.

Example

Check out this tutorial on clustering with the XLSTAT DBSCAN module on the XLSTAT Help Center:

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

References

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.

Classification and regression random forests

Use this method to fit a classification or a regression model on a sample described by qualitative and / or quantitative variables. The method efficiently handles large datasets with a large number of variables.

- **Classification** (qualitative response variable): the model allows to predict the belonging of observations to a class, on the basis of explanatory quantitative and / or qualitative variables.
- **Regression** (continuous response variable): the model allows to build a predictive model for a quantitative response variable based on explanatory quantitative and / or qualitative variables.

In this section:

Description

Random forests are methods that provide predictive models for classification and regression. The method implements binary decision trees, in particular CART trees proposed by Breiman et al. (1984).

The main weakness of CART is its instability which has been studied by Breiman (1994). The remedy, innovative and profoundly statistical, is to exploit the natural variability of the estimation methods by combining two fundamental mechanisms: random perturbation of the trees and combination of a set of trees, rather than selection of a single one of them.

The general idea behind the random forests method is that instead of trying to get a unique optimal tree, we generate several predictors, and then combine their respective predictions.

Two variants are implemented in XLSTAT. Bagging for "Bootstrap aggregating" proposed by Breiman (1996), and Random Input introduced by Breiman in (2001).

The general principle of the method is to aggregate a collection of predictors (here CART trees), to obtain a more efficient final predictor.

CART Tree

The procedure involves partitioning observations by creating the most homogeneous possible groups of observations from the perspective of the variable to predict. Several iterations are necessary : at each iteration we divide the observations into $k = 2$ classes to explain the response variable. The first division is obtained by choosing the explanatory variable which will provide the best separation of the observations on the basis of a quality measure. That division defines subpopulations ("nodes") of the tree. The process is repeated for each subpopulation until no further separation is possible. We then obtain terminal nodes called "leaves" of the tree. Each leaf is characterized by a specific path through the tree, which is called a "rule". The set of rules for all leaves define the model.

Quality measure:

- In the case of regression, the response variable is quantitative. In order to obtain the optimal split, we try to minimize at each node the variance of the child nodes t_L and t_R . The variance of a node t is defined by:

$$\sum_{X_i \in t} (Y_i - \overline{y}(t))^2$$

Where Y_i is the value of the response variable associated to observation i and where $\overline{y}(t)$ is the average of the outputs associated to node t.

- In the case of classification, the response variable Y is qualitative with J categories. The quality measure used for the split in this case is the Gini impurity index. The Gini impurity index of a node t is defined by:

$$i(t) = 1 - \sum_{J} p^2(j|t)$$

with $p(j|t)$ the probability of having the modality j of Y knowing that we are in the node$t$.

In the case of a quantitative explanatory variable, all the possible binary partitions are tested, so we have an infinity of possible tests. Nevertheless, the size n of the learning sample L_n being fixed, we have at most n distinct values for a quantitative variable, therefore at most $n - 1$ associated binary questions.

For a qualitative explanatory variable, each grouping in two groups of k modalities is tested (i.e. $2^k - 1$ possibilities).

After each generation of a new subnode, the stop criteria are checked, and if none of the conditions are fulfilled, the node will be considered as an initial node, and the process is iterated.

Stop criteria:

- Pure node: The node contains only observations of one category or one value of the response variable.
- Variance equals zero: The variance of the response variable associated to observations of a node is null.
- No partitioning allows to improve the quality measure.
- Maximum tree depth: The level of the node has reached the user defined maximum tree depth.
- Minimum size for a parent-node: The node contains fewer observations than the user defined minimum size for a parent-node.
- Minimum size for a son-node: After splitting this node, there is at least one sub-node which size is smaller than the user defined minimum size for a child-node.
- Max nodes: The maximum number of terminal nodes has reached the limit set by the user.

Bagging

The idea here is the following: build CART trees from different bootstrap samples, modify the predictions, and so build a varied collection of predictors. The aggregation step allows then to obtain a robust and more efficient predictor.

Construction:

- We build q samples $L_n^1, \ldots, L_n^q$ by random sampling of n observations among n or sampling of k observations out of n, with k being defined by the user and is strictly lower than n. q corresponds to the number of trees requested by the user.
- For $l = 1, \ldots, q$, we build a CART tree g_l from the sample L_n^l
- We aggregate the predictions of trees $g_1, \ldots, g_q$ by:
- majority vote in classification: $g(X) = max_{j \in 1 \ldots J} \sum_{l=1}^{q} I_{g_l(x)=j}$, j being the class predicted by the tree g_l for any observation x.
- Average of the individual predictions of trees in regression: $\frac{1}{q} \sum_{l=1}^{q} g_l(x)$, $g_l(x)$ being the value predicted by tree g_l for any $x \in X$

Random Input

The Random Input variant is an important modification of the bagging. Its objective is to increase the independent between the models (trees), in order to obtain a final model with better performance. The fundamental difference, here, is that for each sample L_n^q the trees are

not built based on the classical approach of CART but on a different one as follows: to split a node, we randomly select a number m of variables ($m \leq P$, Pbeing the number of explanatory variables), and then look for the best split point according to the m selected variables. The variable selection at each node, is done uniformly and without replacement (each variable has a probability $1/P$ of being chosen).

OOB error measurement:

Let L_n^l be a sample. For each observation (X_i, Y_i), i = 1, ..., n of the learning set, if L_n^l does not contains (X_i, Y_i) we say that the latter is "Out-Of-Bag" (OOB) for that sample.

We will call OOB sample, the sample made up of all the Out-Of-Bag observations for L_n^l

The calculation of OOB error is as follows:

For each observation(X_i, Y_i), i = 1, ..., n of the learning set L_n :

1. Select the sample L_n^l for which (X_i, Y_i) is OOB;
2. Predict the value taken by this observation with all the trees built on these samples;
3. Aggregate the predictions of these trees to make the final prediction$\hat{Y}$ of Y;
4. The error is then calculated:
5. Mean square error in regression:

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

- Proportion of observations misclassified in classification:

$$\frac{1}{n}\sum_{i=1}^{n} I_{(\hat{Y}_i \neq Y_i)}$$

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: click this button to start the calculations.

Cancel: click this button to close the dialog box without doing any calculations.

Help: click this button to display help.

 : click this button to reload the default options.

 : click this button to delete the data selections.

 : click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Response variable: select the dependent variable(s) you want to model. If a column header has been selected, check that the "Variable labels" option has been activated.

Response type: confirm the type of response variable you have selected:

- **Quantitative**: Activate this option if the selected dependent variables are quantitative.
- **Qualitative**: Activate this option if the selected dependent variables are qualitative.

X / Explanatory variables:

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected must be of the numerical type. If a variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If a variable header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (response and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in

the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Options tab:

Forest parameters

Sampling: Select the sampling method.

- **Random with replacement**: Observations are chosen randomly and may occur several times in the sample.
- **Random without replacement**: Observations are chosen randomly and may occur only once in the sample.

Method: Choose the forest type

- Bagging
- Random Input

Sample size: Enter the size k of the sample to generate for the trees construction.

Number of trees: Enter the desired number of trees q in the forest.

Stop conditions:

Construction time (in seconds): Enter the maximum time allowed for the construction of all trees in the forest. Past that time, if the desired number of trees in the forest could not be built, the algorithm stops and returns the results obtained using the trees built until then.

Convergence: Activate this option to check the convergence of the algorithm every X trees (X defined by the user). Once 100 trees are built, we check the OOB error every X trees, and if it has not changed by more than ±3%, the algorithms stops.

Tree parameters:

- **Minimum parent size**: Enter the minimum number of observations that a node must contain to be split.
- **Minimum child size**: Enter the minimum number of observations that every newly created node must contain after a possible split in order to allow the splitting.
- **Maximum depth**: Enter the maximum trees depth.
- **Max nodes**: Activate this option to set the maximum number of terminal nodes a tree can have.
- **Mtry**: Number of variables $m \leq P$ to randomly choose at each node. Note that when $m = P$ we are in the case of bagging.
- **Complexity parameter** (classification only): Enter the value of the complexity parameter (CP). The construction of a tree does not continue unless the overall impurity is reduced

by at least a factor CP. That value must be less than 1.

Validation tab:

Validation: Activate this option if you want to use a sub-sample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: Activate this option if you want to select data to use them in prediction mode. If this option is activated, you need to make sure that the prediction dataset is structured as the learning dataset: same variables with the same order in the selections. If variable labels option is active for the learning data, the first row of the selections listed below must correspond to the variable labels.

Quantitative: Activate this option to select the quantitative explanatory variables.

Qualitative: Activate this option to select the qualitative explanatory variables.

Observations labels: Activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Variable labels: Activate this option if the first row of the data selections (explanatory variables, observations labels) includes a header.

Missing datatab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

OOB predictions: Activate this option to display the vector of Out-Of-Bag predictions.

OOB predictions details: Activate this option to display OOB predictions details

Confusion matrix (classification only): Activate this option to display the table showing the number of well- and badly-classified observations for each of the categories.

Variable importance: Activate this option to display the variable importance measures

- **Normalize**: Activate this option to normalize the importance measures.
- **Standard deviation**: Activate this option to display for each variable the standard deviation of its importance measure.

Charts tab:

OOB error evolution: Activate this option to display the chart showing the evolution of the OOB error according to the number of trees.

Variable importance: Activate this option to display the chart showing the variable importance measures

Predictions chart (régression only) : Activate this option to display the predictions chart: Predictions for the dependent variable versus the dependent variable.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

OOB error:

- **Regression**: Mean square error (based on Out-Of-Bag data).

- **Classification**: Classification error rate (based on Out-Of-Bag data).

Chart of OOB error evolution : This chart shows the OOB error evolution according to the number of trees.

Confusion matrix (classification only): The confusion matrix contains information about the classifications observed and predicted by the algorithm (predictions based on « Out-Of-Bag » data). The performance of the algorithm can be evaluated using this confusion matrix. The diagonal contains the correct predictions. The higher the sum of the elements of the diagonal is, the better the classifier is.

OOB predictions: This table contains different information associated with OOB predictions:

- **Number of times OOB**: Number of times observations are 'out-of-bag' (and thus used in computing OOB error estimate).
- **Y and Prediction(Y)**: initial value of the dependent variable and predicted value (regression also shows the residual value). The prediction associated with each observation is made using the set of trees in which the observation was OOB.
- **OOB predictions details**:
    - **Regression**: Table summarizing for each observation the predictions made by all the trees in the forest.
    - **Classification**: Table with one row for each observation of the learning set and one column for each category of the response variable. It contains for each observation the probability that it has to belong to the different category of the response variable (based on the Out-Of-Bag data).

Predictions (validation sample): Predicted classes or values obtained using the predictive model build on learning data are displayed for the validation sample.

Variable importance:

The importance measure for a given variable is the mean error increase of a tree when the observed values of this variable are randomly exchanged in the OOB samples.

For each tree, the prediction error on the out-of-bag data is computed. Then the same is done after permuting each explanatory variable. The difference between the two is then averaged over all trees, and according to the choice of the user, normalized or not by the standard deviation of the differences.

If the standard deviation of the differences is equal to 0 for a variable, the division is not done.

In classification, in addition to the impact of permutations on the overall error of the forest, we also measure the impact on each of the modalities of the response variable.

Results associated to the predictions: Predicted classes or values obtained using the predictive model build on learning data are displayed for the prediction sample.

Example

Tutorials on how to use Classification and regression random forest are available on the XLSTAT Help Center:

https://help.xlstat.com/6602-random-forest-classification-excel-tutorial

https://help.xlstat.com/6551-random-forest-regression-excel-tutorial

References

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. https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

Association rules

Use this tool to discover association rules for a set of items or objects.

In this section:

Description

In 1994, Rakesh Agrawal and Ramakrishnan Sikrant have proposed the Apriori algorithm to identify associations between items in the form of rules. This algorithm is used when the volume of data to be analyzed is important. As the number of items can be several tens of thousands, combinatorics are such that all the rules can not be studied. It is therefore necessary to limit the search for rules to the most important ones. The quality measurements are probabilistic values ??which limit the combinatorial explosion during the two phases of the algorithm, and allow the sorting of the results.

Definition

Items: Depending on the application field, they can be products, objects, patients, events.

Transaction: Identified with a unique identifier, it is a set of items with a minimum of one item. Items can belong to several transactions.

Itemset: A group of items. Itemsets can be found in one or more transactions.

Support: The probability to find item or itemset X in a transaction. Estimated by the number of times an item or itemset is found across all the available transactions. This value lies between 0 and 1.

Rule: A rule defines a relationship between two itemsets X and Y that have no items in common. X->Y means that if we have X in a transaction, then we can have Y in the same transaction.

Support of a rule: The probability to find items or itemsets X and Y in a transaction. Estimated by the number of times both items or itemsets are found across all the available transactions. This value lies between 0 and 1.

Confidence of a rule: The probability to find item or itemset Y in a transaction, knowing item or itemset X is in the transaction. Estimated by the observed corresponding frequency (number of times X and Y are found across all transactions divided by the number of times X is found). This value lies between 0 and 1.

Lift of a rule: The lift of a rule, which is symmetric (Lift(X->Y)=Lift(Y->X)), is the support of the itemset grouping X and Y, divided by the support of X and the support of Y. This value can be any positive real. A lift greater than 1 implies a positive effect of X on Y (or Y on X) and therefore the significance of the rule. A value of 1 means there is no effect and it is as if the items or itemsets are independent. A lift lower than 1, means there is a negative effect of X on Y or reciprocally. As if they were excluding each other.

Let $I = i_1, \ldots, i_m$ be a set of items. Let $T = t_1, \ldots, t_n$ be a set of transactions, such that t_i is a subset of I.

An association rule R is writen in the following way:

$$R : X \to Y, X \in T, Y \in T, X \cap Y = \emptyset.$$

The support for a subset of I is given by:

$$support(X) = Pr(X).$$

The confidence of a rule (R: X ? Y) is given by:

$$confidence(R) = Pr(Y|X).$$

The lift of a rule (R: X ? Y) is given by:

$$\text{lift}(R) = \frac{support(X \cup Y)}{support(X)support(Y)}.$$

Apriori algorithm

This algorithm involves two steps:

1. Generation of subsets of I with a support greater than a minimum support.

2. Generation of association rules from the subsets of I whose confidence is greater than a fixed minimum confidence.

Hierarchies and multilevel approach

XLSTAT proposes to take into account a hierarchy for grouping the items and study the existing rules at different levels. The proposed method can generate association rules for which the causes or consequences belong either to the same level of the hierarchy or to two different levels.

To simplify the reading of the results, Han and Fu (1999) propose two indexes alpha and beta between 0 and 1, to eliminate redundant and unnecessary rules.

A rule is said to be redundant if it is derived from a rule that is covering it hierarchically: the rule R such that $A_1, ..., A_n \rightarrow B_1, ..., B_m$ is redundant if there is a rule R' such that $A_1^{'}, ...A_n^{'} \rightarrow B_1^{'}, ..., B_m^{'}$ with each $A_i^{'}$ and $B_i^{'}$ $(i = 1...n)$ that are either the same or the parent of an element of A.

R is said to be redundant if its confidence $Conf(R)$ lies in the interval

$$[exp(Conf(R)) - \alpha, exp(Conf(R)) + \alpha]$$

with

$$exp(Conf(R)) = \frac{support(B1)}{support(B'1)} * ... * \frac{support(Bm)}{support(B'm)} * phi(R').$$

A rule is said to be useless if it does not provide more information than a rule with the same consequence and with fewer items as antecedents: Let R be the rule (A, B -> C), and $R^{'}$ the rule (A -> C). R is considered useless if its confidence Conf(R) is in the interval $[Conf(R^{'}) - \beta, Conf(R^{'}) + \beta]$.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

  : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Items: Select a table of items and specify the data format. If column headers were selected, please check the option "Labels included" is activated. The available **data formats** are:

- **Transactional**: Choose this format if your data is in two columns, one indicating the transaction (to be selected in the Transactions field), the other item. Typically with this format, there is a column with the transaction IDs, with for each transaction, as many rows as there are items in the transaction, and a column indicating the items. The transactions can be in the first column and selected in this field.
- **List**: Choose this format if your data include one line per transaction while columns contain the names of the items corresponding to the transaction. The number of items per transaction may vary from one line to another. The number of columns in the selection corresponds to the maximum number of items per transaction.
- **Transactions/Variables**: Choose this format if your data correspond to one line per transaction and to one column per variable. This format is such that all transactions have the same number of items, which is the number of variables, and that items from a given variable cannot be present in the same transaction.
- **Contingency table**: Choose this format if your data include one row per transaction and a column per item, with null values ??if the item is not present and a number greater than 1 if it is present.

You also have the option to select the data in a flat file by clicking the [...] button.

Transactions: Select a column with the transaction IDs for each item. This selection is required if the selected format is "Transactional" and if the array of items has only one column. If the items table has two columns, the first column is considered as corresponding to the transaction.

Target items / Target variables: Activate this option to define one or more items that you want to appear on the right side (the consequence) rules. If the data format is Transactional or List, you can select a list of items that must be in the right part (consequence) of the rules to be generated. If the data format is Transactions/Variables you need to select the variable that will be considered as target. All the rules will have in the right part (consequence) a category of the selected variable. If the data format is Contingency table you can select one or more columns that will be used to identify the target items.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if the first row of the selected data contains a header.

Minimum support: Enter the value of the minimum support (between 0 and 1) so that only rules containing subsets with a support greater than that value are generated.

Minimum confidence: Enter the value of the minimum confidence (between 0 and 1) so that only rules with a confidence greater than that value are generated.

Minimum number of antecedents: Enter the minimum number of antecedents for the rules to generate.

Options tab:

Sort: Select the criterion used to sort the results (confidence, support, lift or nothing).

Multilevel tab:

Use hierarchical information: Activate this option if you want to select and use hierarchical information.

Hierarchy: Select a hierarchical table describing the hierarchy of items and the groups that include them. An item can only belong to a group of higher order. You have the option to select the data from a flat file.

Support for each level: select a table of values ??to assign a different support for each hierarchical level.

Cross-level analysis: Choose this option if you want to generate the rules regardless of their level.

Alpha (redundant rules): Select a value between 0 and 1 to remove redundant rules. Leave 0 if you do not want to use this option.

Beta (useless rules): Select a value between 0 and 1 to remove unnecessary rules. Let 0 if you do not want to use this option.

Outputs tab:

Influence matrix: Activate this option to display the influence matrix calculated from the confidence of the association rules.

Matrix of items: Activate this option to display a table showing the relative importance of combinations of items.

Charts tab:

Influence chart: Select this option to display a 2D graph showing the relative importance of various combinations obtained by the rules of association.

Items charts: This chart represents the relative importance combinations of items.

Results

Association rules: This table displays the association rules obtained by the Apriori algorithm as well as different values for each rule.

Matrix of influence: This table is the crosstab between the antecedents and consequences of rules, with as value, the criterion chosen for sorting the rules (confidence, support, or lift) in the Options tab.

Influence chart: représentation 2D montrant de l'importance relative des règles d'associations.

Items matrix: This chart allows you to view the relative importance of combinations of items. This symmetric table shows the average confidence for each combination of items (row / column and column / row). It is therefore an indicator of the strength of link between the items. It is then used to make a Multidimensional scaling (MDS) to obtain the **Items chart** which is a graphical representation of the table.

Example

A tutorial on association rules is available on XLSTAT Help Center:

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

References

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.

Model performance Indicators

Use the Model Performance Indicators tool to evaluate the performance of your predictive model. Depending on the type of variable of interest (quantitative or qualitative), different indicators are proposed.

In this section:

Description

Introduction

When trying to predict the values of a quantitative Y variable, we talk about **regression**, whereas we talk about **classification** when the Y variable we need to predict is qualitative. XLSTAT offers several regression and classification learning models.

Say we have a variable of interest to predict, and the closer the prediction of the algorithm is to the target variable, the better the model will perform.

It is important to be able to evaluate the performance of a model to measure the risks but also to compare several algorithms and/or models.

The Performance Indicators tool was developed to help us answer this question: How much can I trust a model to predict future events?

Available indicators

XLSTAT provides a large number of indicators to evaluate the performance of a model. The following indicators are available in XLSTAT:

Classification

Notations: TP (True Positives), TN (True Negatives), FP (False Positives) et FN (False Negatives).

- **Accuracy**: The accuracy is the ratio (TP+TN)/(TP+TN+FP+FN).
  The closer it is to 1, better is the test.

- **Precision**: Precision is the ratio TP/(TP + FP).
  It corresponds to the proportion of positive predictions that are actually correct. In other words, a model with an accuracy of 0.8 correctly predicts the positive class in 80% of the cases.

- **Balanced accuracy** (binary case only): Balanced accuracy is an indicator used to evaluate the quality of a binary classifier. It's especially useful when the classes are unbalanced, i.e. one of the two classes appears more often than the other. It is calculated as follows: (Sensitivity + Specificity) / 2.

Sensitivity (equivalent to the **True Positive Rate**): Proportion of positive cases that are detected properly by the test. In other words, the sensitivity measures how the test is effective when used on positive individuals. The test is perfect for positive individuals when sensitivity is 1, equivalent to a random draw when sensitivity is 0.5. If it is below 0.5, the test is counter-performing and it would be useful to reverse the rule so that sensitivity is higher than 0.5 (provided that this does not affect the specificity). The mathematical definition is given by: Sensitivity = TP/(TP + FN).

Specificity (also called **True Negative Rate**): Proportion of negative cases that are correctly detected by the test. In other words, specificity measures how the test is effective when used on negative individuals. The test is perfect for negative individuals when the specificity is 1, and equivalent to a random draw when the specificity is 0.5. If it is below 0.5, the test is counter-performing and it would be useful to reverse the rule so that specificity is higher than 0.5 (provided that this does not affect the sensitivity). The mathematical definition is given by: Specificity = TN/(TN + FP).

- **False Positive Rate** (binary case only): Proportion of negative cases that the test detects as positive (FPR = 1-Specificity).
- **False Negative Rate** (binary case only): Proportion of positive cases that the test detects as negative (FNR = 1-Sensitivity)
- **Correct classification**: Number of well-classified observations.
- **Misclassification**: Number of misclassified observations.
- **Prevalence**: Relative frequency of the event of interest in the total sample (TP+FN)/N.
- **F-measure**: The F-measure also called F-score or F1-score can be interpreted as a weighted average of precision and recall or sensitivity. Its value is between 0 and 1. It is defined by: F-measure = 2 * (Precision * Sensibility) / (Precision + Sensibility).
- **NER** (null error rate): It corresponds to the percentage of error that would be observed if the model always predicted the majority class.
- **Cohen Kappa**: It is useful when we want to study the relationship between the response variable and the predictions. The value of Kappa is between 0 and 1. A value of 1 means that there is a total link between the two variables (perfect classification). It is defined as follows:

$$CohenKappa = (Accuracy - pe)(1 - pe) \text{ with } pe = [(TP + FN) * (TP + FP) + (FP + T$$

- **Cramer's V**: The Cramer's V test compares the degree of linkage between the two variables studied. The closer V is to zero, the less dependent the variables studied are. On the other hand, it will be 1 when the two variables are completely dependent. In binary case (2x2 confusion matrix), it takes a value between -1 and 1. Thus, the closer V is to 1, the stronger the link between the two variables is.

$$ V = \sqrt{ \frac{ \chi ^{2} } {W * (nClass - 1)}}, $$ $nClass$ corresponds to the number of categories of the response variable.

- **MCC** (Matthews correlation coefficient): The Matthews correlation coefficient (MCC) or phi coefficient is used in machine learning as a measure of the quality of binary (two-class) classifications, introduced by biochemist Brian W. Matthews in 1975. The MCC is defined identically to Pearson's phi coefficient.

The coefficient takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. The MCC is in essence a correlation coefficient between the observed and predicted binary classifications; it returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation. (for more information: Matthews correlation coefficient).

It is defined as follows:

$$\frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}.$$

- **Roc curve**: The ROC curve (*Receiver Operating Characteristics*) displays the performance of a model and enables a comparison to be made with other models. The terms used come from signal detection theory. The curve of points (1-specificity, sensitivity) is the ROC curve.

- **AUC**: The area under the curve (AUC) is a synthetic index calculated for ROC curves. The AUC is the probability that a positive event is classified as positive by the test, given all possible values of the test. For an ideal model we have AUC = 1 (above in blue), where for a random pattern we have AUC = 0.5 (above in red). One usually considers that the model is good when the value of the AUC is higher than 0.7. A model that performs well should have an AUC between 0.87 and 0.9. A model with an AUC above 0.9 is excellent.

- **Lift curve**: The Lift curve is the curve that represents the Lift value as a function of the percentage of the population. Lift is the ratio between the proportion of true positives and the proportion of positive predictions. A Lift of 1 means that there is no gain over an algorithm that makes random predictions. Usually, the higher the Lift, the better the model.

- **Cumulative gain curve**: The gain curve represents the sensitivity, or recall, as a function of the percentage of the total population. It allows us to see which portion of the data concentrates the maximum number of positive events.

Regression

Notations:

$$\bar{y} = \frac{1}{W} \sum_{i=1}^{n} w_i y_i$$

W is the sum of the weights.

p^* is the number of variables included in the model.

- **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2.$$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

- **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): The root mean square of the log errors (RMSLE) is the square root of the MSLE.

- **MAPE** (Mean Absolute Percentage Error): MAPE also called MAPD for Mean Absolute Percentage Deviation is defined by:

$$MAPE = \frac{1}{W}\sum_{i=1}^{n} \frac{w_i(|y_i - \hat{y}_i|)}{\max(\epsilon, |y_i|)}; \text{ avec } \epsilon = 10^{-16}.$$

  If the observed values are too small or the errors too large, then the MAPE may be greater than 100%. * **R²**: The determination coefficient for the model. This coefficient, which value is between 0 and 1, is defined by:

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2}.$$

The R² is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer R² is to 1, the better the model is. The problem with the R² is that it does not take into account the number of variables used to fit the model. Its value can be negative and, in this case, it means that the model is not suitable for the data.

- **Adjusted R²**: The adjusted determination coefficient for the model. The adjusted R² can be negative if the R² is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}.$$

  The adjusted R² is a correction to the R² which takes into account the number of variables used in the model.

- **Willmott index (dr)**: Used mainly in hydrological models, the redefined amenity index (Willmott et al., 2012) is calculated as follows:

  $$

  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 }| }, &if \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, &if \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.

  $$

It indicates the sum of the magnitudes of the differences between the model-predicted and observed deviations about the observed mean relative to the sum of the magnitudes of the perfect-model ($\hat{y}_i = y_i$, for all i). Its values are between -1 and 1 and the use of absolute values limits the influence of extreme values.

```
* If $dr$ = 0: The model does not do better than one that predicts the observed

* If $dr$ = 0.5: The sum of the prediction errors is half the sum of the errors

* If $dr$ = -0.5: The sum of the prediction errors is twice the sum of the erro

* If $dr$ is close to -1, this may indicate that the model is inefficient or tha
As the lower limit of dr is approached, interpretations should be made cautious
```

- **Mielke and Berry index**: The index is affected by the MAE and can be used for seasonal cases. It is defined as follows:

$$\rho = 1 - \frac{MAE}{n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}|\hat{y}_j - y_i|}$$

  The denominator represents the average value of the MAE over all $n!$ probable permutations of $\hat{y}_i$ with respect to y_i under the null hypothesis that the n pairs ($\hat{y}_i$ and y_i for $i = 1, ...n$) correspond to a random assignment result.

  Note that this measure is symmetric, i.e. inverting $\hat{y}$ and y leads to the same result. It is bounded by 1.

- **Legates and McCabe's index**: Used mostly in hydrological models, the Legates and McCabe index is recommended when there is seasonality or a difference in mean per period. It is defined as follows:

$$E_1 = 1 - \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{\sum_{i=1}^{n}|y_i - \bar{y}|}$$

  A value of E_1 = 1 indicates a perfect model (no error) while E_1 = 0 indicates a model that is not performing better than a model that predicts the mean observed for any new

observation (baseline). Negative values of E_1 indicate an inefficient model, as they describe a "level of inefficiency" with respect to the reference model.

- **AIC**: Akaike's Information Criterion is defined by:

$$AIC = W \ln(\frac{SSE}{W}) + 2p^*$$

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **AICc**: The corrected Akaike information criterion reduces the probability of choosing a model with too many explanatory variables. It is defined by:

$$ A I Cc = AIC + \frac{2p^2+2p}{n-p-1} $$ This criterion would be more efficient than the AIC when the data set is small and/or has a large number of variables. Specifically, when the ratio $\frac{n}{p}$ is less than 40.

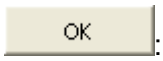
- **SBC**: Schwarz's Bayesian Criterion is defined by:

$$SBC = W \ln(\frac{SSE}{W}) + \ln(W)p^*$$

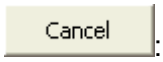
This criterion, proposed by Schwarz (1978) is similar to the AIC, and the aim is to minimize it.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below descriptions of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help menu.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the

arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files  .  : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Y / Response variable: select the response variable(s) you modeled. If a column header has been selected, check that the "Variable labels" option has been activated.

Response type: Confirm the type of dependent variable you have selected:

- **Quantitative**: Choose this option if the selected dependent variable is quantitative.
- **Qualitative**: Choose this option if the selected dependent variable is qualitative.

Predicted values: select the predictions related to the given dependent variable. If a column header has been selected, check that the "Variable labels" option has been activated.

If the response variable is qualitative, only one variable containing the predicted values can be set.

Class probabilities / Scores (classification only): Select this option to indicate, for each class, the probabilities (scores) related to the observations. If column headers have been selected, you must ensure that the labels match the names of the classes to which data are related and that the "Variable labels" option is enabled. In case of binary classification, if only one probability or score column is selected, then high values will be associated with predictions of the positive class.

Explanatory variables (regression only): Select this option if you want to enter the number of explanatory variables included in your model. This information is useful for the calculation of some indicators (adjusted R^2 , AIC, SBC).

Compare to dummy estimator (regression only): select this option if you want to compare the performance of your model with those obtained using a naive regression model. A naive model is a model that predicts the same value for all observations. XLSTAT offers 3 possibilities:

- **Mean**: The comparison is made with a model that predicts the mean of the dependent variable for each observation.
- **Median**: The comparison is made with a model that predicts the median of the dependent variable for each observation.

- **User defined**: The comparison is made with a model that predicts the value given by the user for each observation.

Display results in:

- **New worksheet**: Activate this option to display the results in a new worksheet of the active workbook. In this case, you can give the result sheet a name. If you do not specify a name, a default name will be created.
- **New workbook**: Activate this option to display the results in a new workbook. In this case, you can give the result sheet a name. If you do not specify a name, a default name will be created.
- **Existing cell**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Variable labels: Activate this option if the first row of the data selections includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Outputs tab:

- **Classification**:

  Summary: Activate this option to display the following indicators: Accuracy, Precision, Sensitivity, Specificity, F-measure, False Positive Rate (FPR), False Negative Rate (FNR), number of well classified, number of misclassified.

  Confusion matrix: Activate this option to display the table showing the numbers of well and badly classified observations for each of the categories.

  Prevalence: Activate this option to calculate and display the prevalence.

  F-measure: Activate this option to calculate and display the F-measure.

  Cohen Kappa: Activate this option to calculate and display the Cohen Kappa.

  Nul error rate: Activate this option to calculate and display the Null error rate.

Balanced accuracy: Activate this option to calculate and display the balanced accuracy.

Cramer's V: Activate this option to calculate and display Cramer’s V.

Matthews correlation coefficient: Activate this option to calculate and display Matthews correlation coefficient.

AUC: Activate this option to calculate and display the value of AUC.

- **Regression**:

  Summary: Activate this option to display the following indicators: MAE, MCE, RMCE, MSLE, RMSLE, R^2.

  MAPE: Activate this option to calculate and display the MAPE.

  Willmott's index: Activate this option to calculate and display the value of Willmott's index.

  Legates et McCabe index: Activate this option to calculate and display the value of Legates et McCabe's index.

  Mielke et Berry index: Activate this option to calculate and display the value of de Mielke et Berry index.

  Adjusted R^2: Activate this option to calculate and display the value of adjusted R^2.

  Akaike's AIC: Activate this option to calculate and display the value of AIC.

  Shwarz's SBC: Activate this option to calculate and display the value of SBC.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

Charts tab:

Regression charts:

- **Predictions and residuals**: Activate this option to display the following charts:
  - Response variable versus standardized residuals.
  - Predictions versus standardized residuals.
  - Predictions versus response variable.
  - Bar chart of standardized Residuals.

Classification charts:

- **Roc curve:** Activate this option to display the Roc curve.
- **Lift curve:** Activate this option to display the Lift curve.

- **Cumulative gain curve:** Activate this option to display the cumulative gain curve.

Results

The **predictions and residuals** table shows, for each observation, its weight, the observed value of the dependent variable, the model's prediction, the residual, standardized residuals and Atypical residuals.

The Atypical residuals are identified using the TUCKEY defined box-plot method.

In the TUKEY-defined box plot, the box has the interquartile range (Q3-Q1) as its height, and the whiskers are usually based on 1.5 times the height of the box. In this case, a value is atypical if it is 1.5 times the interquartile range below the 1st quartile or above the 3rd quartile. Based on quartiles, i.e. order statistics, the median and interquartile range are never influenced by extreme values. The value 1.5 is, according to TUKEY, a pragmatic value (rule of thumb) which has a probabilistic reason. If a variable follows a normal distribution, then the area bounded by the box and whiskers should contain 99.3% of the observations. We should therefore find only 0.7% of outliers. If the coefficient is 1, the probability would be 0.957, and it would be 0.999 if the coefficient is 2. For TUKEY the value 1.5 is therefore a compromise to retain as atypical enough observations without retaining too many.

- **minimum and maximum residual(s)**: Marked in green (resp. red), they represent the residuals that have the smallest (resp. largest) deviation from 0. This also allows us to see for each observation which ones are the best (resp. worst) predicted.

Example

An example of how to use the *Model performance Indicators* module is available on the XLSTAT Help Center at the following address:

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

References

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, which stands for "Extreme Gradient Boosting", is a machine learning model that is used for supervised learning problems, in which we use the training data to predict a target/response variable.

XLSTAT provides a no-code user friendly interface to the popular XGBoost open-source library for predictive modeling. XGBoost is a scalable, portable, distributed, open-source C++ library for gradient boosted tree prediction written by the dmlc team; see https://github.com/dmlc/xgboost.

Use this method to fit a classification or regression model on a sample described by qualitative and / or quantitative variables. The method efficiently handles large datasets with a large number of variables.

- **Classification** (qualitative response variable): the model enables to predict the class each observation belongs to, based on explanatory variables that can be quantitative and / or qualitative.
- **Regression** (continuous response variable): the model enables to build a predictive model for a quantitative response variable based on explanatory variables that can be quantitative and / or qualitative.

In this section:

Description

Machine learning models can be fitted to data individually, or combined to other models, creating an ensemble. An ensemble is a combination of simple individual models that together create a more powerful one.

Machine learning boosting is a method that creates such an ensemble. It starts by fitting an initial model (in our case a regression or classification tree) to the data. A second model is then built to focus on accurately predicting the observations that the first model predicted poorly. The combination of these two models is expected to be better than each one of them. This boosting process is then repeated several times, each successive model attempting to correct the shortcomings of the combined boosted ensemble that contains all previous models.

Gradient boosting

Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction

error. The key idea is to set each observation weight for this next model in order to minimize the error. These are calculated the following way:

At each boosting step, for each observation, a score is calculated based on the prediction error of the model.

The name gradient boosting arises from the fact that each weight is set based on the gradient of the error with respect to the prediction. Each new model takes a step in the direction that minimizes the prediction error, in the space of possible predictions for each observation.

Objective Function: Training Loss and Regularization

In supervised learning, the model usually refers to the mathematical structure that enables to predict y_i from the input x_i. The predicted value can have different interpretations, depending on your goal, i.e., regression or classification. For example, it can be transformed by the logit function to get the probability of a class in logistic regression. The parameters are the undetermined part that we need to learn from data. We will use θ to denote the model parameters. The goal is to train the model to find the θ parameters that best fit the training data. In order to train the model, we need to define the objective function to measure how well the model fits the training data.

An important characteristic of objective functions is that they consist of two parts, which are the training loss and the regulation term:

$$Obj(\theta) = L(\theta) + \Omega(\theta)$$

where L is the training loss function and Ω the regalarization term. The training loss measures how predictive our model is with respect to the training data. A common choice for it is the mean squared error (regression task).

The regularization term controls the complexity of the model, which helps us to avoid overfitting. In particular, when observations containing significant errors are present in the input data, increasing the number of iterations may cause a degradation of the overall performance rather than an improvement.

We can thus define the regularized objective function as :

$$Obj(\theta) = \sum_i l(\hat{y}_i, y_i) + \sum_{m=1}^{M} \Omega(\delta_m) \text{with } \Omega(\delta) = \alpha T + \frac{1}{2}\beta \|w\|^2$$

where T is the number of leaves in the tree, M the number of iterations, or boosting steps, that were made. The regularization term Ω is interpreted as a combination of Ridge regularization through the β coefficient and Lasso penalization through the α coefficient. δ_m corresponds to the m-th tree that was built, w corresponds to the vector of weights assigned to it, and w_i represents the score of the i-th leaf.

The general principle is we want both a simple and predictive model. The tradeoff between the two is also referred as bias-variance tradeoff in machine learning.

Shrinkage and subsampling

Besides the regularized objective function, two additional techniques are used to further prevent overfitting. The first technique is shrinkage introduced by Friedman. Shrinkage scales newly added weights by a factor η after each step of tree boosting. Similar to a learning rate in stochastic optimization, shrinkage reduces the influence of each individual tree and leaves space for future trees to improve the model.

The second technique is column (feature) subsampling. This technique is used in **Random Forests** (see the description section for more details). Using column sub-sampling prevents over-fitting even more than the traditional row sub-sampling (which is also supported).

Missing data

The management of missing data is handled in the implementation of XGBOOST. It proposes a default direction for each division if a data is missing. To overcome this shortcoming, the gradient is calculated only on the available values. Note that missing data are allowed only for the input x_i.

Classification table and charts

Among the numerous results provided, XLSTAT can display the classification table (also called confusion matrix) used to calculate the percentage of well- classified observations.

Roc curve: The ROC curve (*Receiver Operating Characteristics*) displays the performance of a model and enables a comparison to be made with other models. The terms used in its name come from signal detection theory. The curve resulting from the points (1-specificity, sensitivity) is the ROC curve.

- **AUC**: The area under the curve (AUC) is a synthetic index calculated for ROC curves. The AUC is the probability that a positive event is classified as positive by the test, given all possible values of the test. For an ideal model we have AUC = 1 (above in blue), where for a random pattern we have AUC = 0.5 (above in red). One usually considers that the model is good when the value of the AUC is higher than 0.7. A model that performs well should have an AUC between 0.87 and 0.9. A model with an AUC above 0.9 is excellent.

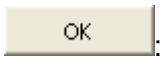
Lift curve: The Lift curve is the curve that represents the Lift value as a function of the percentage of the population. Lift is the ratio between the proportion of true positives and the proportion of positive predictions. A Lift of 1 means that there is no gain over an algorithm that makes random predictions. Usually, the higher the Lift, the better the model.

Cumulative gain curve: The gain curve represents the sensitivity, or recall, as a function of the percentage of the total population. It allows us to see which portion of the data concentrates the maximum number of positive events.

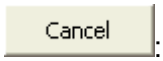
Last, it is suggested to validate the model using a validation sample wherever possible. XLSTAT has several options for automatically generating a validation sample.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: click this button to start the calculations.



: click this button to close the dialog box without doing any calculations.

: click this button to display help.


: click this button to reload the default options.


: click this button to delete the data selections.


: click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Response variable: select the dependent variable you want to model. If a column header has been selected, check that the "Variable labels" option has been activated.

Response type: confirm the type of response variable you have selected:

- **Quantitative**: Activate this option if the selected dependent variables are quantitative.
- **Binary**: Activate this option if the selected dependent variables contain exactly two distinct values.
- **Mutltinomial**: Activate this option if the selected dependent variables has more than two categories.

X / Explanatory variables:

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of the numerical type. If a variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If a variable header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2, …).

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option has been activated.

Options tab:

- **General** tab:

Maximum number of iterations: Enter the maximum number of iterations.

Learning rate: Enter the shrinkage parameter η used after each boosting step to prevent overfitting. Its values can range between [0,1].

Minimum loss reduction: Enter the minimum loss reduction required to make a further partition on a leaf node of a boosting tree. Its values can range between 0 and infinity.

Objective function: Specify the learning objective according to the response variable. The objective options are below: * **Quantitive response**: * **Quadratic**: regression with squared loss. * **Log-quadratic**: regression with squared log loss. * **Logistic**: logistic regression. * **Pseudo-huber**: regression with Pseudo Huber loss, a twice differentiable alternative to absolute loss.

- **Binary response**:
    - **Classification**: logistic regression for binary classification.
- **Multinomial response**:
    - **Classification**: multiclass logistic regression using softmax objective function.

Metric: Specify the evaluation metric according to the learning objective. The choices are listed below:

- **Quantitive response**:
    - **RMSE**: Root Mean Square Error.
    - **RMSLE**: Root Mean Squared Log Error.

    - **MAE**: Mean Absolute Error.
    - **MAPE**: Mean Absolute Percentage Error also called MAPD for Mean Absolute Percentage Deviation.
    - **MPHE**: Mean Pseudo Huber Error.

- **Binary response**:

    - **Error**: Binary classification error rate.
    - **AUCPR**: Area Under the Precision and Recall Curve.

- **Multinomial response**:

    - **Error**: Multiclass classification error rate.
    - **Cross-entropy**: Multiclass log-loss.

Regularization: * **L1 regularization**: Enter the regularization parameter α.

- **L2 regularization**: Enter the regularization parameter β.

Tree parameters * **Minimum son size**: Enter the minimum number of observations that every newly created leaf node must contain after a possible split in order to allow the splitting. * **Maximum depth**: Enter the maximum depth of the trees.

- **Advanced** tab:

  Rows sampling: * **Subsampling ratio**: subsample ratio of the training instance. Setting it to 0.5 means that XGBoost would randomly sample half of the training data prior to growing trees in order to prevent overfitting. Subsampling will occur once in every boosting iteration.

  Column sampling: This is a family of parameters for column subsampling. The parameters can take values in the (0, 1] interval. They have a default value of 1, and specify the fraction of columns to be subsampled.

    - **By tree**: subsample ratio of columns when constructing each tree. Subsampling occurs once for every tree constructed.

    - **By level**: subsample ratio of columns for each level. Subsampling occurs once for every new depth level reached in a tree.

    - **By node**: subsample ratio of columns for each node (split). Subsampling occurs once every time a new split is evaluated. Columns are subsampled from the set of columns chosen for the current level.

Validation tab:

Validation: Activate this option if you want to use a subsample of the data to validate the model.

Validation set: Choose one of the following options to define how to select the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.

- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.
- **Group variable**: if you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: Activate this option if you want to select data to use them in prediction mode. If this option is activated, you need to make sure that the prediction dataset has the same structure as the learning dataset: the same variables, to be selected in the same order. If the "variable labels" option is active for the training data, the first row of the selections listed below must correspond to the variable labels.

Quantitative: Select the quantitative explanatory variables.

Qualitative: Select the qualitative explanatory variables.

Observation labels: Activate this option if observation labels are available. Then select the corresponding data. If this option is not activated, the observation labels are automatically generated by XLSTAT (PredObs1, PredObs2, …).

Variable labels: Activate this option if the first row of the data selections (explanatory variables, observations labels) includes a header.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Ignore missing data: Activate this option to ignore missing data. If missing data are present for the explanatory variable(s) they will be handled by the XGBOOST algorithm using the process described in the description. Missing data are not allowed in the response variable.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

Correlations: Activate this option to display the correlation matrix.

Predictions and residuals(Regression only): Activate this option to display the predictions and residuals for all the observations

Results by object(classification only): Activate this option to display for each observation, the observed category, the predicted category, and the probabilities corresponding to the various categories of the dependent variable.

Statistics for each iteration: Activate this option to display the table showing the evaluation metric evolution across each iteration.

Confusion matrix(classification only): Activate this option to display the table showing the numbers of well and wrongly classified observations for each of the categories.

Variable importance: Activate this option to display the variable (feature) importance measures. XLSTAT gives the importance measures below :

The **Frequency** corresponds to the percentage representing the number of times a feature has been used in trees.

The **Gain** corresponds to the relative contribution of a feature to the model and is calculated by taking the ratio between each feature's total contribution and the total contribution for all the features in the model. The higher the Gain metric is, the more important the predictive feature is.

The **Cover** metric is the proportion of observations that are related to a feature. When a feature is used to split a node that is just before a leaf, we say that the observations in this node are covered by the feature. For example, let's suppose that you have 100 observations, 4 features and 3 trees, and that feature 1 is used to decide the leaf node for 10, 5, and 2 observations in tree1, tree2 and tree3 respectively. The metric will count cover for this feature as 10+5+2 = 17 observations. This will be calculated for all 4 features and the cover will be 17 expressed as a percentage of all features' cover metrics.

The Gain is the most relevant attribute to interpret the relative importance of each feature.

Charts tab:

Statistics for each iteration: Activate this option to display the chart showing the evolution of the evaluation metric across each iteration.

Variable importance: Activate this option to display the chart showing the variable importance measures.

Regression charts: Activate this option to display the following charts:

- Response variable versus standardized residuals.
- Predictions versus standardized residuals.

- Predictions versus response variable.
- Bar chart of standardized Residuals.

Confusion plot(classification only): Activate this option to display the confusion plot which allows a synthetic visualization of the classification table. The numbers can be linked either to the width or the surface of the displayed squares.

Roc curve(classification only): Activate this option to display the ROC curve.

Lift curve(classification only): Activate this option to display the Lift curve.

Cumulative gain curve(classification only): Activate this option to display the cumulative gain curve.

Results

The **predictions and residuals** table shows, for each observation, its weight, the observed value of the dependent variable, the model's prediction, the residual, standardized residuals and atypical residuals.

The Atypical residuals are identified using the TUKEY defined box-plot method.

In the TUKEY-defined box plot, the box has the interquartile range (Q3-Q1) as its height, and the whiskers are usually based on 1.5 times the height of the box. In this case, a value is atypical if it is 1.5 times the interquartile range below the 1st quartile or above the 3rd quartile. Based on quartiles, i.e. order statistics, the median and interquartile range are never influenced by extreme values. The value 1.5 is, according to TUKEY, a pragmatic value (rule of thumb) which has a probabilistic reason. If a variable follows a normal distribution, then the area bounded by the box and whiskers should contain 99.3% of the observations. We should therefore find only 0.7% of outliers. If the coefficient were 1, the probability would be 0.957, and it would be 0.999 if the coefficient were 2. For TUKEY the value 1.5 is therefore a compromise to retain as atypical enough observations without retaining too many.

- **Minimum and maximum residual(s)** : marked in green (resp. red), they represent the residuals that have the smallest (resp. largest) deviation from 0. This also allows us to see for each observation which ones are the best (resp. worst) predicted.

Example

A tutorial on how to use XGBOOST is available on the XLSTAT Help Center: https://www.xlstat.com/demo/gbm_en

References

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).

Correlation/Association tests

Correlation tests

Use this tool to compute different kinds of correlation coefficients, between two or more variables, and to determine if the correlations are significant or not. We propose Pearson, Spearman, Kendall and polychoric correlation coefficients. Several visualizations of the correlation matrices are proposed.

In this section:

Description

Four correlation coefficients are proposed to compute the correlation between a set of quantitative variables, whether continuous, discrete or ordinal.

Pearson correlation coefficient: This coefficient corresponds to the classical linear correlation coefficient. This coefficient is well suited for continuous data. Its value ranges from -1 to 1, and it measures the degree of linear correlation between two variables. Note: the squared Pearson correlation coefficient gives an idea of how much of the variability of a variable is explained by the other variable. The p-values that are computed for each coefficient allow testing the null hypothesis that the coefficients are not significantly different from 0. However, these results should be interpreted with caution, as if two variables are independent, their correlation coefficient is zero, but the reciprocal is not true.

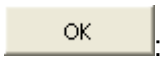
Spearman correlation coefficient (rho): This coefficient is based on the ranks of the observations and not on their value. As for the Pearson correlation, one can interpret this coefficient in terms of explained variability, but here we mean the variability of the ranks.

Kendall correlation coefficient (tau): As for the Spearman coefficient, it is based on ranks. However, this coefficient is conceptually very different. It can be interpreted in terms of probability: it is the difference between the probabilities that the variables vary in the same direction and the probabilities that the variables vary in the opposite direction. When the number of observations is lower than 50 and when there are no ties, XLSTAT gives the exact p-value. If not, an approximation is used. The latter is known as being reliable when there are more than 8 observations.

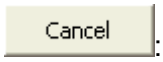
Polychoric correlation coefficient: This coefficient characterizes the relation between two ordinal variables. It is frequently used to analyze survey data with ordinal responses. Under the assumption that the ordinal variables are derived from the discretization of two unobserved quantitative random variables with a normal distribution, the polychoric correlation coefficient aims to measure the relation between those two unobserved quantitative variables.

Dialog box

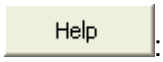
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Observations/variables table: Select a table comprising N observations described by P variables. If column headers have been selected, check that the "Variable labels" option has been activated. For Pearson, Spearman and Kendall, variables must be quantitative, for polychoric variables muste be qualitative ordinal.

Order of categories (only for polychoric correlation):You can select a table that contains the sorted categories for each ordinal variables. This table must have the same number of columns than the observations/variables table. If you do not select this table, the categories are sorted by lexicographical order.

Weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Type of correlation: Choose the type of correlation to use for the computations (see the description section for more details).

Subsamples: Check this option to select a column showing the names or indexes of the subsamples for each of the observations. All computations are then performed subsample by subsample.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (row and column variables, weights) includes a header.

Significance level (%): Enter the significance level for the test of on the correlations (default value: 5%).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Pairwise deletion: Activate this option to remove observations with missing data only when the variables involved in the calculations have missing data. For example, when calculating the correlation between two variables, an observation will only be ignored if the data corresponding to one of the two variables is missing.

Estimate missing data: Activate this option to estimate the missing data before the calculation starts.

- **Mean or mode**: Activate this option to estimate the missing data by using the mean (quantitative variables) or the mode (qualitative variables) for the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data for an observation by searching for the nearest neighbor to the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

Correlations: Activate this option to display the correlation matrix that corresponds to the correlation type selected in the "General" tab.

p-values: Activate this option to display the p-values that correspond to each correlation coefficient. If this option is activated, the correlations that are significant at the selected significance level will be displayed in bold in the correlation matrix.

Coefficients of determination (R2): Activate this option to display the coefficients of determination. These correspond to squared correlations coefficients. When the using the Pearson correlation coefficient, the coefficients of determination are equal to the R^2 of the regression of one variable by the other.

Filter variables on R2: This option allows to filter the display of variables. Different filtering options are available:

- **R2 > (resp. R2 <)**: This option allows to display only variables with at least one determination coefficient (R2) greater (resp. lower) than a given threshold chosen by the user. The threshold must be between 0 and 1.
- **p var with highest (resp.lowest) Sum(R2)**: This option allows to display only the p variables (with p chosen by the user) for which the sum of R2 with other variables is the highest (resp. lowest).

Sort the variables with R2: Activate this option if you want to sort and group variables that are highly correlated. Two sorting method are available:

- **BEA** Method: The BEA (Bond Energy Algorithm) method, initially developed by McCormick (1972), apply a permutation on rows and columns of a square matrix in a way that columns having similar values on rows are close to each other.
- **FPC** Method: The FPC (First Principal Component) method involves doing a PCA on the matrix of coefficients of determination (R2). Then variables are reorganized in such a way that they are sorted in ascending order of their correlations with the first principal component of the PCA.

Charts tab:

Correlation maps: Several visualizations of a correlation matrix are proposed.

- The "**blue-red** " option allows to represent low correlations with cool colors (blue is used for the correlations that are close to -1) and the high correlations are with warm colors (correlations close to 1 are displayed in red color).
- The "**Black and white** " option allows to either display in black the positive correlations and in white the negative correlations (the diagonal of 1s is display in grey color), or to display in black the significant correlations, and in white the correlations that are not significantly different from 0.
- The "**Patterns** " option allows to represent positive correlations by lines that rise from left to right, and the negative correlations by lines that rise from right to left. The higher the absolute value of the correlation, the large the space between the lines.

Scatter plots: Activate this option to display the scatter plots for all two by two combinations of variables.

- **Matrix of plots**: Check this option to display all possible combinations of variables in pairs in the form of a two-entry table with the various variables displayed in rows and in columns.
- **Histograms**: Activate this option so that XLSTAT displays a histogram when the X and Y variables are identical.
- **Q-Q plots**: Activate this option so that XLSTAT displays a Q-Q plot when the X and Y variables are identical.
- **Confidence ellipses**: Activate this option to display confidence ellipses. The confidence ellipses correspond to a confidence interval you can choose for a bivariate normal distribution with the same means and the same covariance matrix as the variables represented in abscissa and ordinates.

Image tab:

Image: If you choose to display the correlation matrix and/or the coefficients of determination matrix in the result sheet, you can activate this option to display the matrices as images. If a sorting and/or a filtering method have been applied, the images will take them into account. This option can be very useful when correlation matrices contain a high number of variables in order to see quickly which variables have the same structure. The following options are available for the produced images:

- **Variable labels**: This option allows to display the variable labels on the top of the image.
- **Grid**:This option allows to display a grid between each cell in order to visually separate variables on the image.
- **Legend**: This option allows to display a legend which indicate at which value correspond each color on the image.

Results

The correlation matrix and the table of the p-values are displayed. The correlation maps allow to identify potential structures in the matrix, of to quickly identify interesting correlations.

Example

A tutorial on how to compute a Spearman correlation coefficient and the corresponding significance test is available on the XLSTAT Help Center:

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

References

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.

RV Coefficient

Use this tool to compute the similarity between two matrices of quantitative variables recorded from the same observations or two configurations resulting from multivariate analyses for the same set of observations..

In this section:

Description

This tool allows computing the RV coefficient between two matrices of quantitative variables recorded from the same observations. The RV coefficient is defined as (Robert and Escoufier, 1976; Schlich, 1996):

$$RV(W_i, W_j) = \frac{trace(W_i, W_j)}{\sqrt{trace(W_i, W_i)trace(W_j, W_j)}}$$

where

$$trace(W_i, W_j) = \sum_{l,m} w_{lm}^i w_{lm}^j$$

is a generalized covariance coefficient between matrices W_i and W_j, and

$$trace(W_i, W_i) = \sum_{l,m} [w_{lm}^i]^2$$

is a generalized variance of matrix W_i and $w_{l,m}^i$ is the (l,m) element of matrix W_i.

The RV coefficient is a generalization of the squared Pearson correlation coefficient. The RV coeffcient is between 0 and 1. The closer to 1 the RV, the more similar the two matrices W_i and W_j.

XLSTAT offers the possibility:

- to compute the RV coefficient between two matrices, including all variables form both matrices;

- to choose the k first variables from both matrices and compute the RV coefficient between the two resulting matrices.

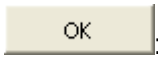
XLSTAT allows testing if the obtained RV coefficient is significantly different from 0 or not.

Two methods to compute the p-value are proposed by XLSTAT. The user can choose between a p-value computed using on an approximation of the exact distribution of the RV statistic with the Pearson type III approximation (Kazi-Aoual et *al.*, 1995), and a p-value computed using Monte Carlo resamplings.

Note: the XLSTAT_RVcoefficient spreadsheet function can be used to compute the RV coefficient between two matrices of quantitative variables.

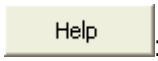
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.

: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Matrix A: Select the data that correspond to N observations described by P quantitative variables. If a column header has been selected, check that the "Column labels" option is activated.

Matrix B: Select the data that correspond to N observations described by Q quantitative variables. If a column header has been selected for Matrix B, a column header should be slected for matrix B.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selections (variables, observations labels) includes a header.

Row labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Column labels" option is activated you need to include a header in the selection.

Options tab:

Selected variables:

All: Choose this option to compute the RV coefficient between Matrix A and Matrix B using all variables from both matrices.

User defined: Choose this option to compute the RV coefficient between sub-matrices of Matrix A and Matrix B with the same number of variables. Then, enter the number of variables to be selected. For example to compute the RV coefficient on the first two variables (or the first two dimensions when comparing results from multidimensional analysis), enter 2 for both **From** and **To**. To compute RV coefficients for a series of number of variables, enter a for **From** and b for **To** where a<b. For example, to compute, RV coefficients for the first 2, 3 and 4 variables, enter 2 for **From** and 4 for **To**.

p-value computation:

Extrapolation: Choose this option to use an approximation of the exact distribution of the RV statistic with the Pearson type III approximation to compute the p-value associated with the RV coefficient.

permutation: Choose this option to calculate the p-value based on Monte Carlo permutations, and select the number of random permutations to perform or the maximum time to spend.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Outputs tab:

RV coefficients: Activate this option to display the RV coefficient(s), standardized RV coefficient(s), and mean(s) and variance(s) of the RV coefficient distribution.

Adjusted RV coefficients: Activate this option to display the adjusted RV coefficient(s).

p-value: Activate this option to display the p-value(s) associated with the RV coefficient(s).

Charts tab:

Chart of RV coefficients: Activate this option to display a bar chart of the RV coefficient(s) (with colors corresponding to the significance of the associated p-value(s) if the option **p-value** has been selected).

Results

RV coefficients: A table including the RV coefficient(s), standardized RV coefficient(s), and mean(s) and variance(s) of the RV coefficient distribution; and the adjusted RV coefficient(s) and p-value(s) if requested by the user.

Bar chart: A bar chart showing the RV coefficient(s) (with color codes to show significance of the associated p-value(s) if requested).

Example

An example showing how to compute the RV coeffcient is available on the XLSTAT Help Center. To download this data, go to:

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

References

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 on contingency tables (chi-square, ...)

Use this tool to study the association between the rows and the columns of a contingency table, and to test the independence between the rows and the columns.

Note: to build a contingency table from two qualitative variables you can use the "Build a contingency table" feature of XLSTAT.

In this section:

Description

Many association measures and several tests have been proposed to evaluate the relationship between the R rows and the C columns of a contingency table.

Some association measures have been specifically developed for the 2×2 tables. Others can only be used with ordinal variables.

XLSTAT always displays all the measures. However, measures that concern ordinal variables should only be interpreted if the variables are ordinal and sorted in increasing order in the contingency table.

Tests of independence between the rows and the columns of a contingency table

- The **Pearson chi-square** statistic allows to test the independence between the rows and the columns of the table, by measuring to which extent the observed table is far (in the chi-square sense) from the expected table computed using the same marginal sums. The statistic writes:

$$\chi^2_P = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(n_{ij}-f_{ij})^2}{f_{ij}}, \quad \text{with } f_{ij} = \frac{n_{i.}n_{.j}}{n}, \quad n = \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}, \quad n_{i.} = \sum_{j=1}^{C} n_{ij}, \quad n_{.j} = \sum_{i=}$$

One shows that this statistic follows a Chi-square distribution with $(R-1)(C-1)$ degrees of freedom. However, this result is asymptotical and, before using the test, it is recommended to make sure that:

- That n is greater or equal to 20;
- That no marginal sum ($n_{i.}$ ou $n_{.j}$) is less than 5;

- That at least 80% of the expected values (f_{ij}) is above 5.
- In the case where $R = 2$ and $C = 2$, a **continuity correction** has been suggested by Yates (1934). The modified statistic writes:

$$\chi^2_Y = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(|n_{ij} - f_{ij}| - 0.5\right)^2}{f_{ij}}$$

- A test based on the likelihood ratio and on the Wilks' G^2 statistic has been developed as an alternative to the Pearson chi-square test. It consists in comparing the likelihood of the observed table to the likelihood of the expected table defined as for the Pearson chi-square test. G^2 is defined by:

$$G^2 = 2\sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}\log(n_{ij}/f_{ij})$$

As for the Pearson statistic, G^2 follows asymptotically a Chi-square distribution with $(R-1)(C-1)$ degrees of freedom.

- The **Fisher's exact test** allows to compute the probability that a table showing a stronger association between the rows and the columns would be observed, the marginal sums being fixed, and under the null hypothesis of independence between rows and columns. In the case of a 2×2 table, the independence is measured through the *odds ratio* (see below for further details) given by $\theta = (n_{11}n_{22})/(n_{12}n_{21})$. The independence is corresponds to the case where $\theta = 1$. There are three possible alternative hypotheses: the two-sided test corresponds to $\theta \neq 1$, the lower one-sided test to $\theta < 1$ and the upper one-sided test to $\theta > 1$.

XLSTAT allows to compute the Fisher's exact two-sided test when $R = 2$ and $C = 2$. The computing method is based on the network algorithm developed by Mehta (1986) and Clarkson (1993). It may fail in some cases. The user is prompted when this happens.

- **Monte Carlotest**: A nonparametric test based on simulations has been developed to test the independence between rows and columns. A number of Monte Carlo simulations defined by the user are performed in order to generate contingency tables that have the same marginal sums as the observed table. The chi-square statistic is computed for each of the simulated tables. The p-value is then determined by suing the distribution obtained from the simulations.

Association measures (1)

A first series of association coefficients between the rows and the columns of a contingency table is proposed:

- The **Pearson's Phi** coefficient allows to measure the association between the rows and the columns of an $R \times C$ table. In the case of a 2×2 table, its value ranges from -1 to 1 and writes:

$$\phi_P = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1.}n_{2.}n_{.1}n_{.2}}}$$

When $R > 2$ and/or $C > 2$, it ranges between 0 and the minimum of the square roots of $R - 1$ and $C - 1$. In that case, the Pearson's Phi writes:

$$\phi_p = \sqrt{\chi^2_P / n}$$

- **Contingency coefficient**: This coefficient, also derived from the Pearson's chi-square statistic, writes:

$$C = \sqrt{\chi^2_P / (\chi^2_P + n)}$$

- **Cramer's V**: This coefficient is also derived from the Pearson chi-square statistic. In the case of 2×2 table, its value has the [-1; 1] range. It writes:

$$V = \phi_p$$

When $R > 2$ and/or $C >$2, it ranges between 0 and 1 and its value is given by:

$$V = \sqrt{\frac{\chi^2_P / n}{\min(R - 1, C - 1)}}$$

The closer V is to 0, the more the rows and the columns are independent.

- **Tschuprow's T**: This coefficient is also derived from the Pearson chi-square statistic. Its value ranges from 0 to 1 and is given by:

$$T = \sqrt{\frac{\chi^2_P / n}{(R - 1, C - 1)}}$$

The closer T is to 0, the more the rows and the columns are independent.

- **Goodman and Kruskal tau (R/C) and (C/R)**: This coefficient, unlike the Pearson coefficient is asymmetric. It allows to measure the degree of dependence of the rows on the columns (R/C) or vice versa (C/R).

- **Cohen's kappa**: This coefficient is computed on RxR tables. It is useful in the case of paired qualitative samples. For example, we ask the same question to the same individuals at two different times. The results are summarized in a contingency table. The Cohen's kappa, which value ranges from 0 to 1, allows to measure to which extent the answer are identical. The close the kappa is to 1, the higher the association between the two variables.

- **Yule's Q**: This coefficient is used on 2x2 tables only. It is computed using the product of the concordant data $(n_{11}n_{22})$ and the product of the discordant data $(n_{12}n_{21})$. It ranges from -1 to 1. A negative value corresponds to a discordance between the two variables, a value close to 0 corresponds to the independence, and a value close to 1 to the

concordance. The Yule's Q is equal to the Goodman and Kruskal Gamma when the latter is computed on a 2×2 table.

- **Yule's Y**: This coefficient is used on 2×2 tables only. It is similar to the Yule's Q and ranges from -1 to 1.

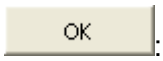
Association measures (2)

A second series of association coefficients between the rows and the columns of a contingency table is proposed. Confidence ranges around the estimated values are available. As the confidence ranges are computed using asymptotical results, their reliability increased with the number of the data.

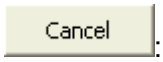
- **Goodman and Kruskal Gamma**: This coefficient allows to measure on a -1 to 1 scale the degree of concordance between two ordinal variables.
- **Kendall's tau**: This coefficient, also referred to as tau-b, allows to measure on a -1 to 1 scale the degree of concordance between two ordinal variables. Unlike the Gamma coefficient, the Kendall's tau allows to take ties into account.
- **Stuart's tau**: This coefficient, also referred to as tau-c, allows to measure on a -1 to 1 scale the degree of concordance between two ordinal variables. As the Kendall's tau, the tau-c allows to take ties into account. In addition, it allows to adjust for the size of the table.
- **Somers' D (R/C) and (C/R)**: This coefficient is an asymmetrical alternative to the Kendall's tau. In the (R/C) case, the rows are assumed to depend on the columns and reciprocally in the (C/R) case; the correction for ties applies only to the "explanatory" variable.
- **Theil's U (R/C) and (C/R)**: The asymmetric coefficient U of uncertainty of Theil (R/C) allows to measure the proportion of uncertainty of the row variable that is explained by the column variable, and reciprocally in the C/R case. These coefficients range from 0 to 1. The symmetric version of the coefficient that ranges from 0 to 1 is computed using the two asymmetric (R/C) and (C/R) coefficients.
- **Odds ratio and Log(Odds ratio)**: The odds ratio is given in the case of 2×2 by $\theta = (n_{11}n_{22})/(n_{12}n_{21})$. θ varies from 0 to infinity. θ can be interpreted as the increase in chances of being in column 1, when being in row 1 compared to when being in row 2. The case $\theta = 1$ corresponds to no advantage. When $\theta > 1$, the probability is θ times higher for row 1 than for row 2. We compute the logarithm of the odds because its variance is easier to compute, and because it is symmetric around 0, which allows to obtain a confidence interval. The confidence of the odds ration itself is computed by taking the exponential of the confidence interval on the log(odds ratio).

Dialog box

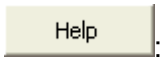
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Contingenc y table: If the data format selected is "contingency table", select the data that correspond to the contingency table. If row and column labels are included, make sure that the "Labels included" option is checked.

Row variable(s): If the data format selected is "contingency table", select the data that correspond to the variable(s) that will be used to construct the rows of the contingency table(s).

Column variable(s): If the data format selected is "contingency table", select the data that correspond to the variable(s) that will be used to construct the columns of the contingency table(s).

By group analysis: Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis on each group separately.

Data format: Select the data format.

- **Contingency table**: Activate this option if your data are correspond to a contingency table.
- **Qualitative variables**: Activate this option if your data are available as two qualitative variables to be used to create a contingency table.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if the row and column labels of the contingency table are selected.

Variable labels: Activate this option if the first row of the data selections (data and observations labels) includes a header.

Options tab:

Chi-square test: Activate this option to display the statistics and the interpretation of the Chi-square test of independence between rows and columns.

Likelihood ratio test: Activate this option to perform the Wilks G^2 likelihood ratio test.

Monte Carlo method: Activate this option to compute the p-value using Monte Carlo simulations.

Significance level (%): Enter the significance level for the test.

Fisher's exact test: Activate this option to compute the Fisher's exact test. In the case of a 2×2 table, you can choose the **alternative hypothesis**. In the other cases, the two-sided is automatically used (see the description section for more details).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Replace missing data by 0: Activate this option if you consider that missing data are equivalent to 0.

Replace missing data by their expected value: Activate this option if you want to replace the missing data by the expected value. The expectation is given by:

$$E(n_{ij}) = \frac{n_{i.}n_{.j}}{n}$$

where $n_{i.}$ is the row sum, $n_{.j}$ is the column sum, and n is the grand total of the table before replacement of the missing data.

Outputs tab:

List of combines: Activate this option to display the table that lists all the possible combines between the two variables that are used to create a contingency table, and the corresponding frequencies.

Contingency table: Activate this option to display the contingency table.

Inertia by cell: Activate this option to display the inertia for each cell of the contingency table.

Chi-square by cell: Activate this option to display the contribution to the chi-square of each cell of the contingency table.

Significance by cell: Activate this option to display a table indicating, for each cell, if the actual value is equal (=), lower (<) or higher (>) than the theoretical value, and to run a test (Fisher's exact test of on a 2x2 table having the same total frequency as the complete table, and the same marginal sums for the cell of interest), in order to determine if the difference with the theoretical value is significant or not.

Association coefficients: Activate this option pour display the various association coefficients.

Observed frequencies: Activate this option to display the table of the observed frequencies. This table is almost identical to the contingency table, except that the marginal sums are also displayed.

Theoretical frequencies: Activate this option to display the table of the theoretical frequencies computed using the marginal sums of the contingency table.

Proportions or percentages / Row: Activate this option to display the table of proportions or percentages computed by dividing the values of the contingency table by the marginal sums of each row.

Proportions or percentages / Column: Activate this option to display the table of proportions or percentages computed by dividing the values of the contingency table by the marginal sums of each column.

Proportions or percentages / Total: Activate this option to display the table of proportions or percentages computed by dividing the values of the contingency table by the sum of all the cells of the contingency table.

Summary across groups: Activate this option to display a summary of all contingency tables.

Charts tab:

3D view of the contingency table: Activate this option to display the 3D bar chart corresponding to the contingency table.

Contingency table: Activate this option to display the contingency table chart.

Proportions or percentages / Row: Activate this option to display the chart related to the *Proportions or percentages / Row* tab.

Proportions or percentages / Column: Activate this option to display the chart related to the *Proportions or percentages / Column* tab.

Summary across groups: Activate this option to display the charts associated with each of the groups in the summary table.

Chart options:

- **Chart type**

- **Grouped**: Choose this option to display the graphs as bars grouped by modality.
- **Stacked bars**: Choose this option to display the chart as stacked bars. These charts are used to compare the frequencies of sub-samples to those of a full sample.

- **Bar charts**
  - **Frequencies**: Choose this option to display the frequencies corresponding to each bar.
  - **Percentages**: Choose this option to display the % of population corresponding to each bar.

Results

The results that are displayed correspond to the various statistics, tests and association coefficients described in the description section.

Example

A tutorial on Chi-square and Fisher's exact tests is available on the XLSTAT Help Center:

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

References

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.

Cochran-Armitage trend test

Use this tool to test if a series of proportions, possibly computed from a contingency table, can be considered as varying linearly with an ordinal or continuous variable.

In this section:

Description

The Cochran-Armitage test allows to test if a series of proportions, can be considered as varying linearly with an ordinal or continuous score variable.

If X is the score variable, the statistic that is computed to test for the linearity is given by:

$$z = \frac{\sum_{i=1}^{r} n_{i1}(X_i - \overline{X})}{\sqrt{p_{+1}(1-p_{+1})s^2}} \quad \text{avec} \quad s^2 = \sum_{i=1}^{r} n_{i+}(X_i - \overline{X})^2$$

Note: if X is an ordinal variable, the minimum value of X has no influence on the value of z.

In the case of the two-tailed (or two-sided) test, the null (H_0) and alternative (H_a) hypotheses are:

- $H_0 : z = 0$
- $H_a : z \neq 0$

Note: z is asymptotically distributed as a standard Normal variable. Some statistical programs use z^2 to test the linearity. z^2 follows a χ^2 distribution with one degree of freedom.

In the one-tailed case, you need to distinguish the left-tailed (or lower- tailed or lower one-sided) test and the right-tailed (or upper-tailed or upper one-sided) test. In the left-tailed test, the following hypotheses are used:

- $H_0 : z = 0$
- $H_a : z < 0$

If H_a is chosen, one concludes that the proportions decrease when the score variable increases.

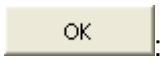
In the right-tailed test, the following hypotheses are used:

- $H_0 : z = 0$
- $H_a : z > 0$

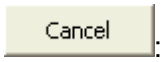
If H_a is chosen, one concludes that the proportions increase when the score variable increases.

Dialog box

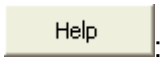
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Contingency table: Select a contingency table. If the column labels of the table have been selected, make sure the "Column labels" option is checked.

Proportions: Select the column (or row if in row mode) that contains the proportions. If a column has been selected, make sure the "Column labels" option is checked.

Sample sizes: If you selected proportions, you must select the corresponding sample sizes. If a column has been selected, make sure the "Column labels" option is checked.

Row labels: Activate this option to select the labels of the rows.

Data format:

- **Contingency table**: Activate this option I your data are contained in a contingency table.
- **Proportions**: Activate this option if your data are available as proportions and sample sizes.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if column headers have been selected within the selections.

Scores: You can choose between ordinal scores (1, 2, 3, ...) or user defined scores.

- **Ordinal**: Activate this option to use ordinal scores.
- **User defined**: Activate this option to select the scores. If a column has been selected, make sure the "Column labels" option is checked.

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Significance level (%): Enter the significance level for the test (default value: 5%).

Asymptotic p-value: Activate this option to compute the p-value based on the asymptotic distribution of the z statistic.

Monte Carlo method: Activate this option to compute the p-value using Monte Carlo simulations. Enter the number of simulations to perform.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics.

Charts tab:

Proportions: Activate this option to display a scatter plot with the scores as abscissa and the proportions as ordinates.

Results

The results include a summary table with the input data, a chart showing the proportions as a function of the scores. The next results correspond to the test itself, and its interpretation.

Example

An example showing how to run a Cochran Armitage trend test is displayed on the XLSTAT Help Center:

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

References

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.

Mantel test

Use this test to compute the linear correlation between two proximity matrices (simple Mantel test), or to compute the linear correlation between two matrices knowing their correlation with a third matrix (partial Mantel test).

In this section:

Description

Mantel (1967) proposed a first statistic to measure the correlation between two proximity (similarity or dissimilarity) and symmetric A and B matrices of size n:

$$z(AB) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij} b_{ij}$$

The standardized Mantel statistic, easier to use because it varies between -1 and 1, is the Pearson correlation coefficient between the two matrices:

$$r(AB) = \frac{1}{n(n-1)/2 - 1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\frac{a_{ij} - \bar{a}}{s_a} \right) \left(\frac{b_{ij} - \bar{b}}{s_b} \right)$$

Notes:

In the case where the similarities or dissimilarities would be ordinal, one can use the Spearman or Kendall correlation coefficients.

In the case where the matrices are not symmetric, the computations are possible.

While it is not a problem to compute the correlation coefficient between two sets of proximity coefficients, testing their significance can not be done using the usual approach that is used to test correlations: to use the latter tests, one needs to assume the independence of the data, which is not the case here. A permutation test has been proposed to determine if the correlation coefficient can be considered as showing a significant correlation between the matrices or not.

In the case of the two-tailed (or two-sided) test, the null (H0) and alternative (Ha) hypotheses are:

- H0 : $r(AB) = 0$
- Ha : $r(AB) \neq 0$

In the one-tailed case, you need to distinguish the left-tailed (or lower- tailed or lower one-sided) test and the right-tailed (or upper-tailed or upper one-sided) test. In the left-tailed test, the following hypotheses are used:

- H0 : $r(AB) = 0$
- Ha : $r(AB) < 0$

In the right-tailed test, the following hypotheses are used:

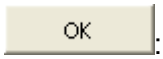
- H0 : $r(AB) = 0$
- Ha : $r(AB) > 0$

The Mantel test consists of computing the correlation coefficient that would be obtained after permuting the rows and columns of one of the matrices. The p-value is calculated using the distribution of the r (AB) coefficients obtained from S permutations. In the case where n, the number of rows and columns of the matrices, is lower than 10, all the possible permutations can easily be computed. If n is greater than 10, one needs to randomly generate a set of S permutations in order to estimated the distribution of r (AB).

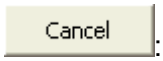
A Mantel test for more than two matrices has been proposed (Smouse *et al.*, 1986): when we have three proximity matrices A, B and C, the partial Mantel statistic r(AB.C) for the A and B matrices knowing the C matrix is computed as a partial correlation coefficient. In order to determine if the coefficient is significantly different from 0, a p-value is computed using random permutations as described by Smouse *et al* (1986).

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Matrix A: Select the first proximity matrix. If the row and column labels are included, make sure the "labels included" option is checked.

Matrix B: Select the second proximity matrix. If the row and column labels are included, make sure the "labels included" option is checked.

Matrix C: Activate this option if you want to compute the partial Mantel test. Then select the third proximity matrix. If the row and column labels are included, make sure the "labels included" option is checked.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if the row and column labels have been selected.

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Significance level (%): Enter the significance level for the test.

Exact p-values: Activate this option so that XLSTAT tries to compute all the possible permutations when possible, to obtain an exact distribution of the Mantel statistic.

Number of permutations: Enter the number of permutations to perform in the case where it is not possible to generate all the possible permutations.

Type of correlation: Select the type of correlation to use to compute the standardized Mantel statistic.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Charts tab:

Scatter plot: Activate this option to display a scatter plot using the values of matrix A on the X axis and the values of the matrix B on the Y axis.

Histogram: Activate this option to display the histogram computed from the distribution of r(AB) based on the permutations.

Results

The displayed results correspond to the standardized Mantel statistic, to the corresponding p-value for the selected alternative hypothesis. A first level interpretation of the test is provided. The histogram of the r(AB) distribution is displayed if the corresponding option has been checked. The observed value of r(AB) is displayed on the histogram.

Example

An example showing how to use the Mantel test is displayed onXLSTAT Help Center:

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

References

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.

Parametric tests

One-sample t and z tests

Use this tool to compare the mean of a normally-distributed sample with a given value.

In this section:

Description

Let the average of a sample be represented by $\hat{\mu}$. To compare this mean with a reference value μ_0, two parametric tests are possible:

- Student's t test if the true variance of the population from which the sample has been extracted is not known; the variance of sample s^2 is used as variance estimator.
- The z test if the true variance σ^2 of the population is known.

These two tests are said to be parametric as their use requires the assumption that the samples are distributed normally. Moreover, it also assumed that the observations are independent and identically distributed. The normality of the distribution can be tested beforehand using the normality tests.

Three types of test are possible depending on the alternative hypothesis chosen:

For the two-tailed test, the null H0 and alternative Ha hypotheses are as follows:

- H0: $\hat{\mu} = \mu_0$
- Ha: $\hat{\mu} \neq \mu_0$

In the left one-tailed test, the following hypotheses are used:

- H0: $\hat{\mu} = \mu_0$

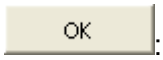
- Ha: $\hat{\mu} < \mu_0$

In the right one-tailed test, the following hypotheses are used:

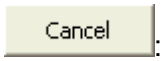
- H0: $\hat{\mu} = \mu_0$
- Ha: $\hat{\mu} > \mu_0$

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data: Select the data in the Excel worksheet.

Data format: choose the data format.

- **One column/row per sample**: Activate this option for XLSTAT to consider that each column (column mode) or row (row mode) corresponds to a sample. You can then test the hypothesis on several samples at the same time.
- **One sample**: Activate this for XLSTAT to consider that all the selected values, whatever the number of rows or columns belong to the same sample.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

z Test: Activate this option to carry out a z test.

Student's t Test: Activate this option to carry out Student's t test.

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Theoretical mean: Enter the value of the theoretical mean with which the mean of the sample is to be compared.

Significance level (%): Enter the significance level for the tests (default value: 5%).

Where a z test has been requested, the population variance value must be entered.

Variance for the z test:

- **Estimated using samples**: Activate this option for XLSTAT to estimate the variance of the population from the sample data. This should, in principle, lead to a t test, but this option is offered for teaching purposes only.
- **User defined**: enter the value of the known variance of the population.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected samples.

Results

The results displayed by XLSTAT relate to the various statistics of the tests selected and the interpretation arising from these.

Example

References

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.

Two-sample t and z tests

Use this tool to compare the means of two normally distributed independent or paired samples.

In this section:

Description

Parametric t and z tests are used to compare the means of two samples. The calculation method differs according to the nature of the samples. A distinction is made between independent samples (for example a comparison of annual sales by shop between two regions for a chain of supermarkets), or paired samples (for example if comparing the annual sales within the same region over two years).

The t and z tests are known as parametric because the assumption is made that the samples are normally distributed. This hypothesis could be tested using normality tests.

Comparison of the means of two independent samples

Take a sample S_1 comprising n_1 observations, of mean $\hat{\mu_1}$ and variance s_1^2. Take a second sample S_2, independent of S_1 comprising n_2 observations, of mean $\hat{\mu_2}$ and variance s_2^2. Let D be the assumed difference between the means (D is 0 when equality is assumed).

As for the z and t tests on a sample, we use:

- Student's t test if the true variance of the populations from which the samples are extracted is not known;
- The z test if the true variance σ^2 of the population is known.

Student's t Test

The use of Student's t test requires a decision to be taken beforehand on whether variances of the samples are to be considered equal or not. XLSTAT gives the option of using Fisher's F test to test the hypothesis of equality of the variances and to use the result of the test in the subsequent calculations.

If we consider that the two samples have the same variance, the common variance is estimated by:

$$s^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)$$

The test statistic is therefore given by:

$$t = \frac{(\hat{\mu_1} - \hat{\mu_2} - D)}{s\sqrt{1/n_1 + 1/n_2}}$$

The t statistic follows a Student distribution with $n_1 + n_2 - 2$ degrees of freedom.

If we consider that the variances are different, the statistic is given by:

$$t = \frac{(\hat{\mu_1} - \hat{\mu_2} - D)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

A change in the number of degrees of freedom was proposed by Satterthwaite:

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Note: when $n_1 = n_2$, we simply have $df = 2(n_1 - 1)$.

Cochran and Cox (1950) proposed an approximation to determine the p-value. It is given as an option in XLSTAT.

z-Test

For the z-test, the variance s^2 of the population is presumed to be known. The user can enter this value or estimate it from the data (this is offered for teaching purposes only). The test statistic is given by:

$$z = \frac{(\hat{\mu_1} - \hat{\mu_2} - D)}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

The z statistic follows a normal distribution.

Comparison of the means of two paired samples

If two samples are paired, they have to be of the same size. Where values are missing from certain observations, either the observation is removed from both samples or the missing values are estimated.

We study the mean of the calculated differences for the n observations. If d is the mean of the differences, s² the variance of the differences and D the supposed difference, the statistic of the t test is given by:

$$t = \frac{(d - D)}{s/\sqrt{n}}$$

The t statistic follows a Student distribution with n-1 degrees of freedom.

For the z test, the statistic is as follows where σ^2 is the variance

$$z = \frac{(d - D)}{\sigma/\sqrt{n}}$$

The z statistic follows a normal distribution.

Alternative hypotheses

Three types of test are possible depending on the alternative hypothesis chosen:

For the two-tailed test, the null H0 and alternative Ha hypotheses are as follows:

- H0 : $\hat{\mu_1} - \hat{\mu_2} = D$
- Ha : $\hat{\mu_1} - \hat{\mu_2} \neq D$

In the left -tailed test, the following hypotheses are used:

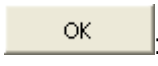
- H0 : $\hat{\mu_1} - \hat{\mu_2} = D$
- Ha : $\hat{\mu_1} - \hat{\mu_2} < D$

In the right-tailed test, the following hypotheses are used:

- H0 : $\hat{\mu_1} - \hat{\mu_2} = D$
- Ha : $\hat{\mu_1} - \hat{\mu_2} > D$

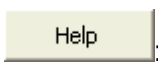
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.

: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data / Sample 1: If the format of the selected data is "one column per variable", select the data for the various samples in the Excel worksheet. If the format of the selected data is "one column per sample" or "paired samples", select a column of data corresponding to the first sample.

Sample identifiers / Sample 2: If the format of the selected data is "one column per variable", select the data identifying the samples to which the selected data values correspond (several columns can be selected). If the format of the selected data is "one column per sample" or "paired samples", select a column of data corresponding to the second sample.

Data format: choose the data format.

- **One column/row per sample**: Activate this option to select one column (or row in row mode) per sample.
- **One column/row per variable**: Activate this option for XLSTAT to carry out as many tests as there are columns/rows, given that each column/row must contain the same number of rows/columns and that a sample identifier which enables each observation to be assigned to a sample must also be selected.
- **Paired samples**: Activate this option to carry out tests on paired samples. You must then select a column (or row in row mode) per sample, all the time ensuring that the samples are of the same size.

Weights: This option is only available if the data format is "**One column/row per variable** " or if the data re paired Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column/rw labels" option is activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

z-test: Activate this option to carry out a z test.

Student's t test: Activate this option to carry out Student's t test.

Options tab:

Alternative hypotheses: Choose the alternative hypothesis to be used for the test (see description).

Hypothesized difference (D): Enter the value of the supposed difference between the samples.

Significance level (%): Enter the significance level for the tests (default value: 5%).

Two methods are available to compute the p-value. Choose among the **asymptotic** or **Monte Carlo** methods. In the case of the Monte Carlo method you can set the maximum time you want to spend computing the p-value.

Where a z test has been requested, the value of the known variance of the populations, or, for a test on paired samples, the variance of the difference must be entered.

Variances for the z test:

Estimated using samples: Activate this option for XLSTAT to estimate the variance of the population from the sample data. This should, in principle, lead to a t test, but this option is offered for teaching purposes only.

User defined: Enter the values of the known variances of the populations.

Sample variances for the t-test:

Assume equality: Activate this option to consider that the variances of the samples are equal.

Cochran-Cox: Activate this option to calculate the p-value by using the Cochran and Cox method where the variances are assumed to be unequal.

Use an F test: Activate this option to use Fisher's F test to determine whether the variances of both samples can be considered to be equal or not.

Multiple samples: If the option *One column/row per variable* is selected, and several columns have been entered in the *Sample identifiers* field, two choices are offered:

- **Merge samples**: Activate this option to merge the columns entered in the *Sample identifiers* field. The tests will be performed according to this new vector.
- **Treat indpendently**: Activate this option to independently replicate the analysis for each of the columns entered in the *Sample identifiers* field.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Ignore missing data: Activate this option to ignore missing data. This option is only available if the format *one column per sample* has been selected.

Remove the observations:

- **For the corresponding sample**: Activate this option to remove the observations with missing data, only for the sample that corresponds to the missing data
- **For all samples**: Activate this option to remove observations with missing data for all the samples.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected samples.

Confidence interval: Activate this option to display the confidence interval on fhe difference between the means.

Detailed results: If the tests are carried out for several variables, activate this option so that the detailed results for each test are displayed.

Summary of comparisons: If the tests are carried out for several variables, activate this option so that the summary table of all comparisons is displayed.

If we have several sample identifiers and if all of them are binary (only two groups), the summary also includes the associated p-value chart.

Charts tab:

Dominance diagram: Activate this option to display a dominance diagram in order to make a visual comparison of the samples.

Distributions: Activate this option to display the distribution of the decision variable.

Results

The results displayed by XLSTAT relate to the various statistics of the tests selected and the interpretation arising from these.

The dominance diagram enables a visual comparison of the samples to be made. The first sample is represented on the x-axis and the second on the y-axis. To build this diagram, the data from the samples is sorted first of all. When an observation in the second sample is greater than an observation in the first sample, a "+" is displayed. When an observation in the second sample is less than an observation in the first sample, a "-" is displayed. In the case of a tie, a "o" is displayed.

Example

An example showing how to run a two sample Student's t test is available at:

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

References

Cochran W.G. and Cox G.M. (1950). Experimental Designs. John Wiley & 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.

Comparison of the means of k samples

If you want to compare the means of k samples, you have to use the ANOVA tool that allows to run many post hoc comparison tests.

One sample variance test

Use this tool to compare the variance of a normally-distributed sample with a given value.

In this section:

Description

Let us consider a samle of n independent normally distributed observations. One shows that the sample variance, s² follows a Chi-squared distribution with $n-1$ degrees of freedom.

$$s^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$$

where σ^2 is the theoretical sample variance. This allows us to compute a confidence interval around the variance.

To compare this variance to a reference value σ_0^2, a parametric test is proposed. It is based on the following statistic:

$$\chi_0^2 = (n-1)\frac{s^2}{\sigma_0^2}$$

which follows a Chi-squared distribution with $n-1$ degrees of freedom.

This test is said to be parametric as its use requires the assumption that the samples are distributed normally. Moreover, it also assumed that the observations are independent and identically distributed. The normality of the distribution can be tested beforehand using a normality test.

Three types of test are possible depending on the alternative hypothesis chosen:

For the two-tailed test, the null H0 and alternative Ha hypotheses are as follows:

- H0 : $s^2 = \sigma_0^2$
- Ha : $s^2 \neq \sigma_0^2$

In the left one-tailed test, the following hypotheses are used:

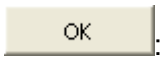
- H0 : $s^2 = \sigma_0^2$
- Ha : $s^2 < \sigma_0^2$

In the right one-tailed test, the following hypotheses are used:

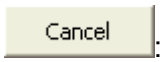
- H0 : $s^2 = \sigma_0^2$
- Ha : $s^2 > \sigma_0^2$

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data: Select the data in the Excel worksheet.

Data format: choose the data format.

- **One column/row per sample**: Activate this option for XLSTAT to consider that each column (column mode) or row (row mode) corresponds to a sample. You can then test the hypothesis on several samples at the same time.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Theoretical variance: Enter the value of the theoretical mean with which the mean of the sample is to be compared.

Significance level (%): Enter the significance level for the tests (default value: 5%).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected samples.

Confidence interval: Activate this option to display the confidence intervals.

Detailed results: If the tests are carried out for several variables, activate this option so that the detailed results for each test are displayed.

Summary of comparisons: If the tests are carried out for several variables, activate this option so that the summary table of all comparisons is displayed.

Charts tab:

Distributions: Activate this option to display the distribution of the decision variable.

Results

The results displayed by XLSTAT relate to the confidence interval around the variance and to the test comparing the observed variance to the theoretical variance.

Example

An example showing how to run a one sample variance test is available at:

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

References

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.

Two-sample comparison of variances

Use this tool to compare the variances of two samples.

In this section:

Description

Three parametric tests are offered for the comparison of the variances of two samples. Take a sample S_1 comprising n_1 observations with variance s_1^2. Take a second sample S_2 comprising n_2 observations with variance $s_1 22$. XLSTAT offers three tests for comparing the variances of the two samples.

Fisher's F test

Let R be the assumed ratio of the variances (R is 1 when equality is assumed).

The test statistic F is given by:

$$F = \frac{s_1^2}{Rs_2^2}$$

This statistic follows a Fisher distribution with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom if both samples follow a normal distribution.

Three types of test are possible depending on the alternative hypothesis chosen:

For the two-tailed test, the null H0 and alternative Ha hypotheses are as follows:

- H0: $s_1^2 = s_2^2 R$
- Ha: $s_1^2 \neq s_2^2 R$

In the left-tailed test, the following hypotheses are used:

- H0: $s_1^2 = s_2^2 R$

- Ha: $s_1^2 < s_2^2 R$

In the right-tailed test, the following hypotheses are used:

- H0: $s_1^2 = s_2^2 R$
- Ha: $s_1^2 > s_2^2 R$

Levene's test

Levene's test can be used to compare two or more variances. It is a two-tailed test for which the null and alternative hypotheses are given by the following for the case where two variances are being compared:

- H0: $s_1^2 = s_2^2$
- Ha: $s_1^2 \neq s_2^2$

The statistic from this test is more complex than that from the Fisher test and involves absolute deviations at the mean (original article by Levene, 1960) or at the median (Brown and Forsythe, 1974). The use of the mean is recommended for symmetrical distributions with averagely thick tails. The use of the median is recommended for asymmetric distributions.

The Levene statistic follows a Fisher's F distribution with 1 and $n_1 + n_2 - 2$ degrees of freedom.

Bartlett's homogeneity of variances test

Bartlett's test can be used to compare two or more variances. This test is sensitive to the normality of the data. In other words, if the hypothesis of normality of the data seems fragile, it is better to use Levene's or Fisher's test. On the other hand, Bartlett's test is more powerful if the samples follow a normal distribution.

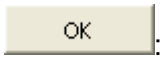
This also is a two-tailed test which can be used with two or more variances. Where two variances are compared, the hypotheses are:

- H0: $s_1^2 = s_2^2$
- Ha: $s_1^2 \neq s_2^2$

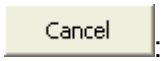
Bartlett's statistic follows a Chi-square distribution with one degree of freedom.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data / Sample 1: If the format of the selected data is "one column per variable", select the data for the various samples in the Excel worksheet. If the format of the selected data is "one column per sample", select a column of data corresponding to the first sample.

Sample identifiers / Sample 2: If the format of the selected data is "one column per variable", select the data identifying the two samples to which the selected data values correspond. If the format of the selected data is "one column per sample", select a column of data corresponding to the second sample.

Data format: choose the data format.

- **One column/row per sample**: Activate this option to select one column (or row in row mode) per sample.
- **One column/row per variable**: Activate this option for XLSTAT to carry out as many tests as there are columns/rows, given that each column/row must contain the same number of rows/columns and that a sample identifier which enables each observation to be assigned to a sample must also be selected.

Column/row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

Fisher's F test: Activate this option to use Fisher's F test (see description).

Levene's test: Activate this option to use Levene's test (see description).

- **Mean**: Activate this option to use Levene's test based on the mean.
- **Median**: Activate this option to use Levene's test based on the median.

Bartlett's test: Activate this option to use Bartlett's test (see description).

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Hypothesized ratio (R): Enter the value of the supposed ratio between the variances of the samples.

Significance level (%): Enter the significance level for the tests (default value: 5%).

Two methods are available to compute the p-value. Choose among the **asymptotic** or **Monte Carlo** methods. In the case of the Monte Carlo method you can set the maximum time you want to spend computing the p-value.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected samples.

Confidence interval: Activate this option to display the confidence intervals.

Detailed results: If the tests are carried out for several variables, activate this option so that the detailed results for each test are displayed.

Summary of comparisons: If the tests are carried out for several variables, activate this option so that the summary table of all comparisons is displayed.

Charts tab:

Distributions: Activate this option to display the distribution of the decision variable.

Results

The results displayed by XLSTAT relate to the various statistics of the tests selected and the interpretation arising from these.

Example

An example showing how to run a Fisher's test do compare the variances of two samples is available at:

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

References

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. & Rohlf F.J. (1995). Biometry. The Principles and Practice of Statistics in Biological Research. Third Edition. Freeman, New York.

k-sample comparison of variances

Use this tool to compare the variances of k samples.

In this section:

Description

Two parametric tests are offered for the comparison of the variances of k samples ($k = 2$). Take k samples $S_1, S_2, \ldots, S_k$, comprising $n_1^2 = n_2^2 = \ldots = n_k^2$ observations with variances $s_1^2 = s_2^2 = \ldots = s_k^2$.

Levene's test

Levene's test can be used to compare two or more variances. This is a two-tailed test for which the null and alternative hypotheses are:

- H0 : $s_1^2 = s_2^2 = \ldots = s_k^2$
- Ha: There is at least one pair (i, j) such that $s_i^2 \neq s_j^2$

The statistic from this test involves absolute deviations at the mean (original article by Levene, 1960) or at the median (Brown and Forsythe, 1974). The use of the mean is recommended for symmetrical distributions with averagely thick tails. The use of the median is recommended for asymmetric distributions.

The Levene statistic follows a Fisher distribution with $k - 1$ and $n1 + n2 - 2$ degrees of freedom.

Bartlett’s homogeneity of variances test

Bartlett's test can be used to compare two or more variances. This test is sensitive to the normality of the data. In other words, if the hypothesis of normality of the data seems fragile, it is better to use Levene's or Fisher's test. On the other hand, Bartlett's test is more powerful if the samples follow a normal distribution.

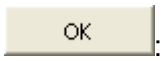
This also is a two-tailed test which can be used with two or more variances. Where two variances are compared, the hypotheses are:

- H0 : $s_1^2 = s_2^2 = \ldots = s_k^2$
- Ha: There is at least one pair (i, j) such that $s_i^2 \neq s_j^2$

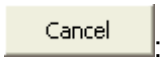
Bartlett's statistic follows a Chi-squared distribution with $k - 1$ degrees of freedom.

Dialog box

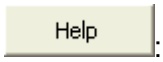
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data / Sample 1: If the format of the selected data is "one column per variable", select the data for the various samples in the Excel worksheet. If the format of the selected data is "one column per sample", select a column of data corresponding to the first sample.

Sample identifiers / Sample 2: If the format of the selected data is "one column per variable", select the data identifying the k samples to which the selected data values correspond. If the format of the selected data is "one column per sample", select a column of data corresponding to the second sample.

Data format: choose the data format.

- **One column/row per sample**: Activate this option to select one column (or row in row mode) per sample.
- **One column/row per variable**: Activate this option for XLSTAT to carry out as many tests as there are columns/rows, given that each column/row must contain the same number of rows/columns and that a sample identifier which enables each observation to be assigned to a sample must also be selected.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

Levene's test: Activate this option to use Levene's test (see description).

- **Mean**: Activate this option to use Levene's test based on the mean.
- **Median**: Activate this option to use Levene's test based on the median.

Bartlett's test: Activate this option to use Bartlett's test (see description).

Options tab:

Significance level (%): Enter the significance level for the tests (default value: 5%).

Two methods are available to compute the p-value. Choose among the **asymptotic** or **Monte Carlo** methods. In the case of the Monte Carlo method you can set the maximum time you want to spend computing the p-value.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected samples.

Results

The results displayed by XLSTAT relate to the various statistics of the tests selected and the interpretation arising from these.

Example

References

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.

Multidimensional tests (Mahalanobis, ...)

Use this tool to compare two or more samples simultaneously on several variables.

In this section:

Description

The tests implemented in this tool are used to compare samples described by several variables. For example, instead of comparing the average of two samples as with the Student t test, we compare here simultaneously for the same samples averages measured for several variables.

Compared to a procedure that would involve as many Student t tests as there are variables, the method proposed here has the advantage of using the structure of covariance of the variables and of obtaining an overall conclusion. It may be that two samples are different for a variable with a Student t test, but that overall it is impossible to reject the hypothesis that they are similar.

Mahalanobis distance

The Mahalanobis distance, from the name if the Indian statistician Prasanta Chandra Mahalanobis (1893-1972), allows computing the distance between two points in a p-dimensional space, while taking into account the covariance structure across the p dimensions. The square of the Mahalanobis distance writes:

$$d_M^2 = (\vec{x}_1 - \vec{x}_2)' \Sigma^{-1} (\vec{x}_1 - \vec{x}_2)$$

In other words, it is the transposed of the vector of the difference of coordinates for p dimensions between the two points, multiplied by the inverse of the covariance matrix multiplied by the vector of differences. The Euclidean distance corresponds to the Mahalanobis distance where the covariance matrix is the identity matrix, which means that the variables are standardized and independent.

The Mahalanobis distance can be used to compare two groups (or samples) because the Hotelling T^2 statistic defined by:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} d_M^2$$

follows a Hotelling distribution, if the samples are normally distributed for all variables. The F statistic that is used for the comparison test where the null hypothesis H0 is that the means of the two samples are equal, is defined by:

$$F = \frac{n_1 + n_2 - (p+1)}{(n_1 + n_2 - 2)p} T^2$$

This statistic follows a Fisher's F distribution with p and $n1 + n2 - p - 1$ degrees of freedom if the samples are normally distributed for all the variables.

Note: This test can only be used if we assume that the samples are normally distributed and have identical covariance matrices. The second hypothesis can be tested with the Box or Kullback tests available in this tool.

If we want to compare more than two samples, the test based on the Mahalanobis distance can be used to identify possible sources of the difference observed at the global level. It is then recommended to use the Bonferroni correction for the alpha significance level. For k samples, we use the following significance level should be used:

$$\alpha^* = \frac{2\alpha}{k(k-1)}$$

Wilks' lambda

The Wilks' lambda statistic follows the three parameters Wilks' distribution defined by:

$$\Lambda(p, m, n) = \frac{|A|}{|A + B|}$$

where A and B are two semi-defined positive matrices that respectively follow Wishart $W_p(I, m)$ and $W_p(I, n)$ distributions, where I is the identity matrix.

When we want to compare the means of p variables for k independent groups (or samples or classes), testing as null hypothesis H0 that the p averages are equal, if we assume that the covariance matrices are the same for the k groups, is equivalent to calculate the following statistic:

$$\Lambda(p, n - k, k - 1) = \frac{|W|}{|W + B|}$$

where

- W is the pooled within-group covariance matrix,

- B is the pooled between-groups covariance matrix,

- n is the total number of observations.

The distribution of the Wilks lambda is complex, so we use instead the Rao's F statistic given by:

$$F = \frac{(1-\Lambda^{1/s})}{\Lambda^{1/s}} \frac{m_2}{m_1}$$

with

$$s = \sqrt{\frac{p^2(k-1)^2-4}{p^2+(k-1)^2-5}}$$
$$m_1 = p(k-1)$$
$$m_1 = s[n-(p+k+2)/2] - p(k-1)/2 + 1$$

One can show that if the sample size is large, then F follows a Fisher's F distribution with $m1$ and $m2$ degrees of freedom. When $p \leq 2$ or $k = 2$, the F statistic is exactly distributed as $F(m1, m2)$.

Notes:

- This test can only be used if we assume that the p variables are normally distributed and have identical covariance matrices.
- The second hypothesis that covariances matrices are the same for the k groups can be tested with the Box or Kullback tests available in this tool.

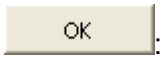
Testing the equality of the within-groups covariance matrices

Box test: The Box test is used to test the assumption of equality for intra-class covariance matrices. Two approximations are available, one based on the Chi-square distribution, and the other on the Fisher distribution.

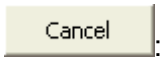
Kullback's test: The Kullback's test is used to test the assumption of equality for intra-class covariance matrices. The statistic calculated is approximately distributed according to a Chi-square distribution.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Observations/variables table: Select a table comprising N objects described by P descriptors. If column headers have been selected, check that the "Variable labels" option has been activated.

Groups: Check this option to select the values which correspond to the identifier of the group to which each observation belongs.

Weights: Activate this option if the observations are weighted. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column labels" option is activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections includes a header.

Options tab:

Wilks' Lambda test: Activate this option to compute the Wilks' lambda test.

Mahalanobis test: Activate this option to compute the Mahalanobis distances as well as the corresponding F statistics and p-values.

- **Bonferroni correction**: Activate this option if you want to use a Bonferroni correction during the computation of the p-values corresponding to the Mahalanobis distances.

Box test: Activate this option to compute the Box test using the two available approximations.

Kullback's test: Activate this option to compute the Kullback's test.

Significance level (%): Enter the significance level for the tests (default value: 5%).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

Correlations: Activate this option to display the correlation matrix.

Covariance matrices: Activate this option to display the inter-class, intra-class, intra-class total, and total covariance matrices.

Results

The results displayed by XLSTAT correspond to the various tests that have been selected.

Example

An example showing how to compare multidimensional samples is available onXLSTAT Help Center:

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

References

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.

z-test for one proportion

Use this test to compare a proportion calculated from a sample with a given proportion.

In this section:

Description

Let n be the number of observations verifying a certain property among a sample of size N. The proportion of the sample verifying the property is defined by $p = n/N$. Let p_0 be a known proportion with which we wish to compare p. Let D be the assumed difference (exact, minimum or maximum) between the two proportions p and p_0. D is usually 0.

The two-tailed (or two-sided) test corresponds to testing the difference between $p - p_0$ and D, using the null (H0) and alternative (Ha) hypotheses shown below:

- H0: $p - p_0 = D$
- Ha: $p - p_0 \neq D$

In the one-tailed case, you need to distinguish the left-tailed (or lower- tailed or lower one-sided) test and the right-tailed (or right-sided or upper one-sided) test. In the left-tailed test, the following hypotheses are used:

- H0: $p - p_0 = D$
- Ha: $p - p_0 < D$

In the right-tailed test the following hypotheses are used:

- H0: $p - p_0 = D$
- Ha: $p - p_0 > D$

This z-test is based on the following assumptions:

- The observations are mutually independent,

- The probability p of having the property in question is identical for all observations,
- The number of observations is large enough, and the proportions are neither too close to 0 nor to 1.

Note: to determine whether N is sufficiently large one should make sure that:

$$\begin{cases} 0 < p - 2\sqrt{p(1-p)/N} \\ p + 2\sqrt{p(1-p)/N} < 1 \end{cases}$$

z statistic

One can find several ways to compute the z statistic in the statistical literature. The most used version is:

$$z = \frac{p - p_0 - D}{\sigma}$$

The large sample approximation leads to the following estimate for its standard deviation *s* :

$$\hat{\sigma^2}(z) = \sqrt{\frac{p(1-p)}{N}}$$

However if one think that proportion we are comparing our sample proportion with might be a better estimate, one can use.

$$\hat{\sigma^2}(\pi) = \sqrt{\frac{p_0(1-p_0)}{N}}$$

This version of the statistic should not be used when *D* is not null.

The z statistic is asymptotically normally distributed. The larger N, the better the approximation. The p-value is computed using the normal approximation.

Confidence intervals

Several methods exist to compute confidence intervals on a proportion. XLSTAT offers the choice between four different versions: Wald, Wilson score, Clopper-Pearson, Agresti Coull.

Dialog box

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

General tab:

Frequency / Proportion: Enter the number of observations n for which the property is observed (see description), or the corresponding proportion (see "data format" below).

Sample size: Enter the number of observations in the sample.

Test proportion: Enter the value of the test proportion with which the proportion observed is to be compared.

Data format: Choose here if you would prefer to enter the value of the **number of observations** for which the property is observed, or the **proportion** observed.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

z-Test: Activate this option to use a z-test.

Options tab:

Alternativehypotheses: Choose the alternative hypothesis to be used for the test (see description).

Hypothesized difference (D): Enter the value of the supposed difference between the proportions.

Significance level (%): Enter the significance level for the test (default value: 5%).

Variance: Select the method used to estimate the variance of the proportion (used only for the confidence interval with the Wald interval).

- **Sample**: Activate this option to compute the variance using the proportion obtained for the sample.
- **Test proportion**: Activate this option to compute the variance using the test proportion and the size of the sample.

Confidence interval: Select the method used to compute the confidence interval (Wald, Wilson score, Clopper-Pearson, Agresti Coull).

Results

The results displayed by XLSTAT relate to the various statistics of the tests selected and the interpretation arising from these.

Example

An example showing how to compare proportions is available on XLSTAT Help Center:

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

References

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.

z-test for two proportions

Use this tool to compare two proportions calculated for two samples.

In this section:

Description

Let n_1 be the number of observations verifying a certain property for sample S_1 of size N_1, and n_2 the number of observations verifying the same property for sample S2 of size $N2$. The proportion of sample S_1 verifying the property is defined by $p_1 = n_1/N_1$, and the proportion for S2 is defined by $p_2 = n_2/N_2$. Let D be the assumed difference (exact, minimum or maximum) between the two proportions p_1 and p_2. D is usually 0.

The two-tailed (or two-sided) test corresponds to testing the difference between $p_1 - p_2$ and D , using the null (H0) and alternative (Ha) hypotheses shown below:

- H0: $p_1 - p_2 = D$
- Ha: $p_1 - p_2 \neq D$

In the one-tailed case, you need to distinguish the left-tailed (or lower- tailed or lower one-sided) test and the right-tailed (or right-sided or upper one-sided) test. In the left-tailed test, the following hypotheses are used:

- H0: $p_1 - p_2 = D$
- Ha: $p_1 - p_2 <> D$

In the right-tailed test the following hypotheses are used:

- H0: $p_1 - p_2 = D$
- Ha: $p_1 - p_2 > D$

This test is based on the following assumptions:

- the observations are mutually independent,

- the probability p_1 of having the property in question is identical for all observations in sample S_1,
- the probability p_2 of having the property in question is identical for all observations in sample S_2,
- the number of observations N_1 and N_2 are large enough, and the proportions are neither too close to 0 nor to 1.

Note: to determine whether N_1 and N_2 are sufficiently large one should make sure that:

$$\begin{cases} 0 < p_1 - 2\sqrt{p_1(1-p_1)/N_1} \\ p_1 + 2\sqrt{p_1(1-p_1)/N_1} < 1 \end{cases}$$

and

$$\begin{cases} 0 < p_2 - 2\sqrt{p_2(1-p_2)/N_2} \\ p_2 + 2\sqrt{p_2(1-p_2)/N_2} < 1 \end{cases}$$

Dialog box

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

General tab:

Frequency 1 / Proportion 1: Enter the number of observations n1 for which the property is observed (see description), or the corresponding proportion (see "data format" below).

Sample size 1: Enter the number of observations in sample 1.

Frequency 2 / Proportion 2: Enter the number of observations n2 for which the property is observed (see description), or the corresponding proportion (see "data format" below).

Sample size 2: Enter the number of observations in sample 2.

Data format: Choose here if you would prefer to enter the values of the **number of observations** for which the property is observed, or the **proportions** observed.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

z-Test: Activate this option to use a z-test.

Monte Carlo method: Activate this option to compute the p-value using Monte Carlo simulations. Enter the number of simulations to perform.

Options tab:

Alternativehypotheses: Choose the alternative hypothesis to be used for the test (see description).

Hypothesized difference (D): Enter the value of the supposed difference between the proportions.

Significance level (%): Enter the significance level for the test (default value: 5%).

Variance: Select the method used to estimate the variance of the difference between the proportions.

- **p1q1/n1+p2q2/n2**: Activate this option to compute the variance using this formula.
- **pq(1/n1+1/n2)**: Activate this option to compute the variance using this formula.

Results

The results displayed by XLSTAT relate to the various statistics of the tests selected and the interpretation arising from these.

Example

An example showing how to compare two proportions is available on XLSTAT Help Center:

http://www.xlstat.com/demo-prop2.htm

References

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.

Comparison of k proportions

Use this tool to compare k proportions, and to determine if they can be considered as equal, or if at least one pair of proportions shows a significant difference.

In this section:

Description

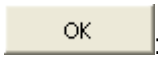
XLSTAT offers three different approaches to compare proportions and to determine whether they can be considered as equal (null hypothesis H0) or if at least two proportions are significantly different (alternative hypothesis Ha):

Chi-square test: This test is identical to that used for the contingency tables;

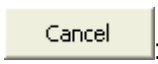
Monte Carlo method: The Monte Carlo method is used to calculate a distribution of the Chi-square distance based on simulations with the constraint of complying with the total number of observations for the k groups. This results in an empirical distribution which gives a more reliable critical value (on condition that the number of simulations is large) than that given by the Chi-square theoretical distribution which corresponds to the asymptotic case.

Marascuilo procedure: It is advised to use the Marascuilo procedure only if the Chi-square test or the equivalent test based on Monte Carlo simulations reject H0. The Marascuilo procedure compares all pairs of proportions, which enables the proportions possibly responsible for rejecting H0 to be identified.

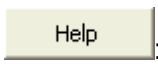
Dialog box



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.

 : Click this button to delete the data selections.

Frequencies / Proportions: Select the data in the Excel worksheet.

Sample sizes: Select the data corresponding to the sizes of the samples.

Sample labels: Activate this option if sample labels are available. Then select the corresponding data. If the "Column labels" option is activated you need to include a header in the selection. If this option is not activated, the row labels are automatically generated by XLSTAT (Sample1, Sample2 …).

Data format: Choose here if you would prefer to enter the value of the **number of observations** for which the property is observed, or the **proportions** observed.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first line of the data selected (frequencies/proportions, sample size and sample labels) contain a label.

Chi-square test: Activate this option to use the Chi-square test.

Monte Carlo method: Activate this option to use the simulation method and enter the number of simulations.

Marascuilo procedure: Activate this option to use the Marascuilo procedure.

Significance level (%): Enter the significance level for the three tests (default value: 5%).

Results

The results of the Chi-square test are displayed first if the corresponding option has been activated. For the Chi-square test and the Monte Carlo method, the p-value is compared with the significance level in order to validate the null hypothesis.

The results obtained from Monte Carlo simulations are all the more close to the Chi-square results the higher the total number of observations and number of simulations. The difference relates to the critical value and the p-value.

The Marascuilo procedure identifies which proportions are responsible for rejecting the null hypothesis. It is possible to identify which pairs of proportions are significantly different by looking at the results in the "Significant" column.

Note: it might be that the Marascuilo procedure does not identify significant differences among the pairs of proportions, while the Chi-square test rejects the null hypothesis. In general, this happens when the two proportions are significantly different as identified by the Marascuilo procedure. More in- depth analysis might be necessary before making a decision.

Example

An example showing how to compare k proportions is available on XLSTAT Help Center:

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

References

Agresti A. (1990). Categorical Data Analysis. John Wiley & Sons, New York.

Marascuilo L. A. and Serlin R. C. (1988). Statistical Methods for the Social and Behavioral Sciences. Freeman, New York.

Multinomial goodness of fit test

Use this tool to check whether the observed frequencies of the values (categories) of a qualitative variable correspond to the expected frequencies or proportions.

In this section:

Description

The multinomial goodness of fit test allows verifying whether the distribution of a sample corresponding to a qualitative variable (or discretized quantitative variable) is consistent with what is expected. The test is based on the multinomial distribution which is the extension of the binomial distribution when there are more than two possible outcomes.

Let k be the number of possible values (categories) for variable X. We write $p_1, p_2, \ldots, p_k$ the probabilities (or densities) corresponding to each value.

Let $n_1, n_2, \ldots, n_k$ be the frequencies of each value for a sample.

The null hypothesis of the test writes:

- H0: The distribution of the values in the sample is consistent with what is expected, meaning the distribution of the sample is not different from the distribution of X.

The alternative hypothesis of the test writes:

- Ha: The distribution of the values in the sample is not consistent with what is expected, meaning the distribution of the sample is different from the distribution of X.

Several methods and statistics have been proposed for this test. XLSTAT offers the following two possibilities:

1. Chi-square test:

We compute the following statistic:

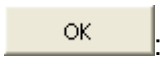
$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i}$$

This statistic is asymptotically distributed as Chi-square with $k - 1 degrees$ of freedom.

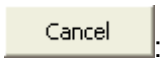
2. Monte Carlo test:

This version of the test overcomes some heavy calculations of the exact method based on the multinomial distribution, and avoids the approximation by the Chi-square distribution that may be of poor quality with small samples. This test consists of a random resampling of N observations in a distribution having the expected properties. For each resampling, we compute the χ^2 statistic, then once the resampling process is finished, we evaluate how many times the value observed on the sample is exceeded, from what we deduce the p-value.

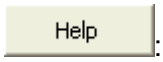
Dialog box



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

Frequencies: Select the data corresponding to the observed frequencies in the Excel worksheet.

Expected frequencies / Expected proportions: Select the data corresponding to the expected frequencies or to the expected proportions. If you select expected frequencies, they must sum to the same value as the sum of the observed frequencies.

Data format: Choose here if you would prefer to select **expected frequencies** or **expected proportions**.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first line of the data selected (frequencies/proportions, sample size and sample labels) contain a label.

Chi-square test: Activate this option to use the Chi-square test.

Monte Carlo method: Activate this option to use the simulation method and enter the number of simulations.

Significance level (%): Enter the significance level for the two tests (default value: 5%).

Results

The results of the Chi-square test are displayed first if the corresponding option has been activated. For the Chi-square test and the Monte Carlo method, the p-value is compared with the significance level in order to validate the null hypothesis.

The results obtained from Monte Carlo simulations are all the more close to the Chi-square results the higher the total number of observations and number of simulations. The difference relates to the critical value and the p-value.

For the Monte Carlo test, a confidence interval on the p-value is displayed.

Example

An example showing how to run a multinomial goodness of fit test is available on XLSTAT Help Center:

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

References

Read T.R.C. and Cressie N.A.C. (1988). Goodness-of-Fit Statistics for Discrete Multivariate Data. Springer-Verlag, New York.

Equivalence test (TOST)

Use this tool to test the equivalence of the two normally distributed independent samples.

In this section:

Description

Unlike classical hypothesis testing, equivalence tests are used to validate the fact that a difference is in a given interval.

This type of test is used primarily to validate bioequivalence. When we want to show the equivalence of two drugs, classical hypothesis testing does not apply, we will use equivalence testing which will validate the equivalence between the two drugs.

In a classical hypothesis test, we try to reject the null hypothesis of equality. As part of an equivalence test, we try to validate the equivalence between two samples. The TOST (two one-sided test) is a test of equivalence that is based on the classical t test used to test the hypothesis of equality between two means.

So we will have two samples, a theoretical difference between the means as well as a range within which we can say that the sample means are equivalent.

The test is known as parametric because the assumption is made that the samples are normally distributed. This hypothesis could be tested using normality tests.

The TOST test uses Student's test to check the equivalence between the means of two samples. A detailed description of such tests can be found in the chapter dedicated to t tests.

XLSTAT offers two equivalent methods to test equivalence using the TOST test.

- Using the 100 * (1-2 * alpha)% confidence interval around the mean. By comparing this interval to the user-defined interval of equivalence, we can conclude the equivalence or non equivalence. Thus, if the confidence interval is within the interval defined by the user, we conclude the equivalence between the two samples. If one of the bounds of the confidence interval is outside the interval defined by the user, then the two samples are not equivalent.

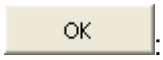
- Using two one-sided tests, one on the right and one on the left. We apply a right one-sided t-test on the lower bound of the interval defined by the user and a left one-sided t-test on the

upper bound of the interval defined by the user. We obtain p-values for both tests. We take the greatest of these p-values as p-value of the equivalence test.

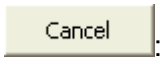
These two tests are similar and should give similar results. They were introduced by Schuirman's (1987).

Dialog box

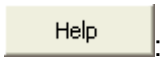
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Sample 1: Select a column of data corresponding to the first sample.

Sample 2: Select a column of data corresponding to the second sample.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

Options tab:

Hypothesized difference (D): Enter the value of the supposed difference between the samples.

Lower bound: Enter the value of the supposed lower bound for equivalence testing.

Upper bound: Enter the value of the supposed upper bound for equivalence testing.

Significance level (%): Enter the significance level for the tests (default value: 5%).

Weights: This option is only available if the data format is "**One column/row per variable** " or if the data re paired Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column/rw labels" option is activated.

Sample variances for the t-test:

Assume equality: Activate this option to consider that the variances of the samples are equal.

Cochran-Cox: Activate this option to calculate the p-value by using the Cochran and Cox method where the variances are assumed to be unequal.

Use an F test: Activate this option to use Fisher's F test to determine whether the variances of both samples can be considered to be equal or not.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected samples.

Results

The first table displays the descriptive statistics associated with the two samples.

The following table of results can be used to validate the hypothesis of equivalence for two means. If the confidence interval around the difference with a confidence level of (1-2 *alpha)*100% is included in the interval defined by the user in the dialog box, then the samples

are equivalent. You have to check if the four values in this table are ordered increasingly. The last line gives an interpretation (equivalence or non equivalence).

The following table allows you to view two one-sided tests based on the bounds defined by the user. The p-value for the test of equivalence is the largest p-value obtained with the one-sided t tests.

Example

An example showing how to run an equivalence test for two samples is available at:

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

References

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.

Nonparametric tests

Comparison of two distributions (Kolmogorov-Smirnov)

Use this tool to compare the distributions of two samples and to determine whether they can be considered identical.

In this section:

Description

The Kolmogorov-Smirnov test compares two distributions. This test is used for distribution fitting tests for comparing an empirical distribution determined from a sample with a known distribution. It can also be used for comparing two empirical distributions.

Note: this test enables the similarity of the distributions to be tested at the same time as their shape and position.

Take a sample S_1 comprising n_1 observations, with F_1 the corresponding empirical cumulative distribution function. Take a second sample S_2 comprising n_2 observations, with F_2 the corresponding empirical cumulative distribution function.

The null hypothesis of the Kolmogorov-Smirnov test is defined by:

H0: $F_1(x) = F_2(x)$

The Kolmogorov statistic is given by:

$$D_1 = \sup_x |F_1(x) - F_2(x)|$$

D_1 is the maximum absolute difference between the two empirical distributions. Its value therefore lies between 0 (distributions perfectly identical) and 1 (separations perfectly separated). The alternative hypothesis associated with this statistic is:

Ha: $F_1(x) \neq F_2(x)$

The Smirnov statistics are defined by:

$$D_2 = \sup_x \{F_1(x) - F_2(x)\}$$

$$D_3 = \sup_x \{F_2(x) - F_1(x)\}$$

The alternative hypothesis associated with D_2 is:

Ha: $F_1(x) < F_2(x)$

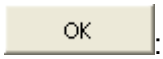
The alternative hypothesis associated with D_3 is:

Ha: $F_1(x) > F_2(x)$

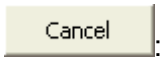
Nikoforov (1994) proposed an exact test method for the Kolmogorov-Smirnov on two samples. This method is used by XLSTAT for the three alternative hypotheses. XLSTAT also enables the supposed difference D between the distributions to be introduced. The value must be between 0 and 1.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data / Sample 1: If the format of the selected data is "one column per variable", select the data for the various samples in the Excel worksheet. If the format of the selected data is "one column per sample", select a column of data corresponding to the first sample.

Sample identifiers / Sample 2: If the format of the selected data is "one column per variable", select the data identifying the two samples to which the selected data values correspond. If the format of the selected data is "one column per sample", select a column of data corresponding to the second sample.

Data format: choose the data format.

- **One column/row per sample**: Activate this option to select one column (or row in row mode) per sample.
- **One column/row per variable**: Activate this option for XLSTAT to carry out as many tests as there are columns/rows, given that each column/row must contain the same number of rows/columns and that a sample identifier which enables each observation to be assigned to a sample must also be selected.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

Kolmogorov-Smirnov test: Activate this option to run the Kolmogorov-Smirnov test (see description).

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Hypothesized difference (D): Enter the value of the maximum supposed difference between the empirical distribution functions of the samples. The value must be between 0 and 1.

Significance level (%): Enter the significance level for the test (default value: 5%).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected samples.

Charts tab:

Dominance diagram: Activate this option to display a dominance diagram in order to make a visual comparison of the samples.

Cumulative histograms: Activate this option to display the chart showing the empirical distribution functions for the samples.

Results

The results displayed by XLSTAT relate to the various statistics of the tests selected and the interpretation arising from these.

References

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.statistics*, 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.

Example

Median test (Mood test)

Use this tool to test if k independent samples have the same median.

In this section:

Description

Proposed in 1950, the Mood test -also called median test- allows the median equality between k independent samples to be test. This test is based on a nonparametric procedure – thus not making any assumption on the distribution of the measurements- and can be considered as a special case of the Pearson Khi² test.

Mood test

Assume that Mi is the median of the i-th sample. The null hypothesis H0 and the alternative Ha associated to the Mood test are:

- H0 : M1 = M2 = … = Mk
- Ha : There is at least one pair (i,j) such as Mi $\neq$ Mj

The U statistic of the Mood test is obtained via the following contingency table (2 x k):

Sample	1	2	…	k	Total
$>$**Median**	O_{11}	O_{12}	…	O_{1k}	a
$\leq$**Median**	O_{21}	O_{22}	…	O_{2k}	b
Total	n_1	n_2	…	n_k	N

If a large number of ties with the median is detected, XLSTAT will automatically count observations equal to the median with those above so that the following statitic remains computable.

We have,

$U = \frac{N^2}{ab} \sum_{i=1}^{k} \frac{(O_{1i} - \frac{n_i a}{N})}{n_i}$,

where N, n, a, b, and O are defined in the contingency table.

This statistic has the property to be asymptotically distributed according to a Khi² distribution at a k-1 degree of freedom. When the number of samples k is 2, Yates (1934) proposed a continuity correction for U. Denote by U_Y this statistic.

$$U_Y = \frac{N(|O_{11}O_{22} - O_{12}O_{21}| - N/2)^2}{n_1 n_2 ab}$$

Computation of the p-values

To compute the p-values corresponding to the various statistics, XLSTAT offers several alternatives:

- Asymptotic method: The p-value is obtained using the asymptotic approximation of the distribution of the U. The reliability of the approximation depends on the number of observations.
- Monte Carlo method: The computation of the p-value is based on random resamplings. The user must set the number of resamplings. A confidence interval on the p-value is provided. The more resamplings are performed, the better the estimation of the p-value.
- Exact method: If the number of samples is 2, the probability to obtain any contingency table setting is determined by the hypergeometric distribution. This approach is often employed when the samples size are small, i.e. when the Khi² approximation is not suitable.

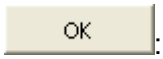
In order to avoid freezing Excel because of too long computations, it is possible with the two latter methods to set the maximum time that should be spent computing the p-value.

Multiple pairwise comparisons

If the p-value is such that the H0 hypothesis has to be rejected, then at least one variable has a different median from the others. To identify which variables are responsible for rejecting H0, a **multiple comparison** procedure can be used.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Observations/ Variables table: Select a table where each row (or column if in column mode) corresponds to an observation, and each column (or row in row mode) corresponds to a variable. If headers have been selected with the data, make sure the "Column labels" option is checked.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if headers have been selected with the input data.

Multiple pairwise comparisons: Activate this option to compute multiple pairwise comparisons.

Options tab:

Significance level (%): Enter the significance level for the test (default value: 5%).

To compute the p-value, you can either choose the asymptotical approximation method, the exact approach or the Monte Carlo resamplings based method (see the description section). In the latter case, you can set the number of resamplings you want to make, and the maximum time you want XLSTAT to spend on making the resamplings.

In some cases, instead of calculating the classical U statistic of the Mood test, the correction for continuity proposed by Yates (see the description section) can be taken into account.

Outputs tab:

Descriptive statistics: Activate this option to compute and display the statistics that correspond to each variable.

Results

Descriptive statistics: This table displays the descriptive statistics that correspond to the k samples.

Mood test: The results that correspond to the Mood test are then displayed, followed by a short interpretation of the test. Results of multiple comparisons are then displayed to identify the variables responsible for rejecting the null hypothesis, if it has been rejected.

Example

An example showing how to run a Mood test is available on XLSTAT Help Center:

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

References

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

One sample Wilcoxon Signed-Rank test

Use this tool to test the null hypothesis that the location parameter (median) of a sample is equal to a given value.

In this section:

Description

This test is the non parametric version of the one sample t test. It is based on ranks and because of that, the location parameter is not here the mean but the median. You should use this test as soon as you have doubts that the normality assumption necessary to apply the t test does not hold. This test allows to test the null hypothesis that the sample median is equal to a given value provided by the user.

If m is assumed to be the median of the sample, there are three possible alternative hypotheses to the null hypothesis:

For the two-tailed test, the null H_0 and alternative H_a hypotheses are as follows:

- $H_0 : Sample\ median = m$
- $H_a : Sample\ median \neq m$

In the left-tailed test, the following hypotheses are used:

- $H_0 : Sample\ median = m$
- $H_a : Sample\ median < m$

In the right-tailed test, the following hypotheses are used:

- $H_0 : Sample\ median = m$
- $H_a : Sample\ median > m$

Wilcoxon signed-rank test

Wilcoxon proposed a test which takes into account the size of the difference within pairs. This test is called the Wilcoxon signed rank test, as the sign of the differences is also involved.

For each observation in the sample $(X_1, X_2, \ldots, X_n)$, we compute its difference with a given median m. The differences are then ordered and the corresponding ranks are computed, and then signed depending on whether they correspond to positive or negative differences. Let $S_1, S_2, \ldots, S_p$ be the positive signed ranks.

The statistic used to test whether the sample median is equal to m is:

$$V_s = \sum_{i=1}^{p} S_i$$

The expectation and the variance of V_s are:

$$E(V_s) = \frac{n(n+1)}{4}$$

and

$$V(V_s) = \frac{n(n+1)(2n+1)}{24}$$

Where they might be ties among the differences, or null differences for certain pairs, we have:

$$E(V_s) = \frac{n(n+1) - d_0(d_0+1)}{4}$$

$$V(V_s) = \frac{[n(n+1)(2n+1) - d_0(d_0+1)(2d_0+1)]}{24} - \frac{\sum_{i=1}^{nd}(d_i^3 - d_i)}{48}$$

where d_0 is the number of null differences, nd the number of distinct differences, and d_i the number of values corresponding to the i'th distinct difference value (it is the same as considering that the d_i's are the number of ties for the i'th distinct difference value).

Where there are no null differences or ties among the differences, if $n \leq 100$, XLSTAT calculates an exact p-value (Lehmann, 1975). Where there are ties, a normal approximation is used. We have:

$$P(V_s \leq \nu) \approx \Phi\left(\frac{\nu - E(V_s) + c}{\sqrt{V(V_s)}}\right)$$

where f is the distribution function for the standardized normal distribution, and c is a continuity correction used to increase the quality of the approximation (c is $\frac{1}{2}$ or $-\frac{1}{2}$ depending on the nature of the test). The approximation is more reliable the higher n is.

The median of the sample and its confidence interval are also computed on the basis of these computations. This estimator is reliable for samples with a symmetric distribution. In the case of

ties the Hodges and Lehmann approach (1963) appears to be still reliable based on Monte Carlo simulations we have made.

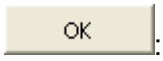
Computation of the p-values

To compute the p-values corresponding to the various statistics, XLSTAT offers several alternatives:

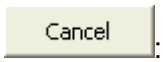
- Asymptotic method: The p-value is obtained using the asymptotic approximation of the distribution of the statistic. The reliability of the approximation depends on the number of samples and on the number of measures per sample.
- Exact method: The computation of the p-value is based on the exact distribution of the statistic.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Samples: Select the data for the various samples in the Excel worksheet.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Theoreticalmedian (D): Enter the value of the supposed median of the samples.

Significance level (%): Enter the significance level for the test (default value: 5%).

Depending on the test that is being used, several methods can be available to compute the **p-value**. Choose among the **asymptotic** and **exact**.

Continuity correction: Activate this option if you want XLSTAT to use the continuity correction when computing the asymptotic p-value (see description).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected samples.

Detailed results: Activate this option to display the detailed results of the tests.

Summary table: In case several samples have been selected activate this option to display a table summarizing the results of the tests.

Results

The results displayed by XLSTAT relate to the various statistics of the tests selected and the interpretation arising from these.

Example

A tutorial showing how to use the Wilcoxon signed-rank test is available on XLSTAT Help Center:

http://www.xlstat.com/demo-1-sample-wilcoxon.htm

References

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.

Comparison of two samples (Wilcoxon, Mann-Whitney, ...)

Use this tool to compare two samples described by ordinal or discrete quantitative data whether independent or paired.

In this section:

Description

To get round the assumption that a sample is normally distributed required for using the parametric tests (z test, Student's t test, Fisher's F test, Levene's test and Bartlett's test), non-parametric tests have been put forward.

As for parametric tests, a distinction is made between independent samples (for example a comparison of annual sales by shop between two regions for a chain of supermarkets), or paired samples (for example if comparing the annual sales within the same region over two years).

If we designate D to be the assumed difference in position between the samples (in general we test for equality, and D is therefore 0), and $P_1 - P_2$ to be the difference of position between the samples, three tests are possible depending on the alternative hypothesis chosen:

For the two-tailed test, the null H_0 and alternative H_a hypotheses are as follows:

- $H_0 : P_1 - P_2 = D$
- $H_a : P_1 - P_2 \neq D$

In the left-tailed test, the following hypotheses are used:

- $H_0 : P_1 - P_2 = D$
- $H_a : P_1 - P_2 < D$

In the right-tailed test, the following hypotheses are used:

- $H_0 : P_1 - P_2 = D$

- $H_a : P_1 - P_2 > D$

Comparison of two independent samples

Three researchers, Mann, Whitney, and Wilcoxon, separately perfected a very similar non-parametric test which can determine if the samples may be considered identical or not on the basis of their ranks. This test is often called the **Mann-Whitney** test, sometimes the Wilcoxon-Mann-Whitney test or the *Wilcoxon Rank-Sum test* (Lehmann, 1975).

We sometimes read that this test can determine if the samples come from identical populations or distributions. This is completely untrue. It can only be used to study the relative positions of the samples. For example, if we generate a sample of 500 observations taken from an $\mathcal{N}(0,1)$ distribution and a sample from a distribution of 500 observations from an $\mathcal{N}(0,4)$ distribution, the Mann-Whitney test will find no difference between the samples.

Let S_1 be a sample made up of n_1 observations $(x_1, x_2, \ldots, x_{n1})$ and S_2 a second sample made up of n_2 observations $(y_1, y_2, \ldots, y_{n1})$ independent of S_1. Let N be the sum of n_1 and n_2.

To calculate the Wilcoxon W_s statistic which measures the difference in position between the first sample S_1 and sample S_2 from which D has been subtracted, we combine the values obtained for both samples, then put them in order. The W_s statistic is the sum of the ranks of one of the samples. For XLSTAT, the sum is calculated on the first sample.

For the expectation and variance of W_s we therefore have:

$$E(W_S) = \frac{1}{2}n_1(N+1) \quad \text{et} \quad V(W_S) = \frac{1}{12}n_1n_2(N+1)$$

The Mann-Whitney U statistic is the sum of the number of pairs (x_i, y_i) where $x_i > y_i$, from among all the possible pairs. We show that

$$E(U) = \frac{n_1n_2}{2} \text{ et} \quad V(U) = \frac{1}{12}n_1n_2(N+1)$$

We may observe that the variances of W_s and U are identical. In fact, the relationship between U and W_s is:

$$W_S = U + \frac{n_1(n_1+1)}{2}$$

The results offered by XLSTAT are those relating to Mann-Whitney's U statistic.

When there are ties between the values in the two samples, the rank assigned to the tied values is the mean of their rank before processing, for example, for two samples of respective size 3 and 3, if the ordered list of values is {1, 1.2, 1.2, 1.4, 1.5, 1.5}, the ranks are initially {1, 2, 3, 4, 5, 6} then after inclusion {1, 2.5, 2.5, 4, 5.5, 5.5}. Although this does not change the expectation of W_s and U, the variance is, on the other hand, modified.

$$V(W_S) = V(U) = \frac{1}{12} n_1 n_2 (N+1) - \frac{n_1 n_2 \sum_{i=1}^{nd} (d_i^3 - d_i)}{12N(N-1)}$$

where nd is the number of distinct values and d_i the number of observations for each of the values.

For the calculation of the p-values associated with the statistic, XLSTAT can use an exact method if the user wants for the following cases:

$U * n_1 * n_2 \leq 10e7$ if there are no ties

$U * nd \leq 5000$ if there are ties.

The calculations may be appreciably slowed down where there are ties. A normal approximation has been proposed to get round this problem. We have:

$$P(U \leq u) \approx \Phi \left(\frac{u - E(U) + c}{\sqrt{V(U)}} \right)$$

where F is the distribution function for the standardized normal distribution, and c is a continuity correction used to increase the quality of the approximation (c is $\frac{1}{2}$ or $-\frac{1}{2}$ depending on the nature of the test). The approximation is more reliable the higher n_1 and n_2 are.

If the user requests that an exact test be used and this is not possible because of the constraints given below, XLSTAT indicates in the results report that an approximation has been used.

A Monte Carlo approximation of the p-value is also possible for this test.

Comparison of two paired samples

Two tests have been proposed for the cases where samples are paired: the **sign test** and the **Wilcoxon signed rank test.**

Let S_1 be a sample made up of n observations $(x_1, x_2, \ldots, x_n)$ and S_2 a second sample paired with S_1, also comprising n observations $(y_1, y_2, \ldots, y_n)$. Let $(p_1, p_2, \ldots, p_n)$ be the n pairs of values (x_i, y_i).

Sign test

Let $N+$ be the number of pairs where $y_i > x_i$, N_0 the number of pairs where $y_i = x_i$, and $N-$ the number of pairs where $y_i < x_i$. We can show that $N+$ follows a binomial distribution with parameters $(n - N_0)$ and probability $\frac{1}{2}$. The expectation and the variance of $N+$ are therefore:

$$E(N+) = \frac{n - N_0}{2} \quad \text{and} \quad V(N+) = \frac{n - N_0}{4}$$

The p-value associated with $N+$ and the type of test chosen (two-tailed, right or left one-tailed) can therefore be determined exactly.

Note: This test is called the sign test as it constructs the differences within the n pairs from the sign. This test is therefore used to compare evolutions evaluated on an ordinal scale. For example, this test would be used to determine if the effect of a medicine is positive from a survey where the patient simply declares if he feels less well, not better, or better after taking it.

The disadvantage of the sign test is that it does not take into account the size of the difference between each pair, data which is often available.

Wilcoxon signed-rank test

Wilcoxon proposed a test which takes into account the size of the difference within pairs. This test is called the *Wilcoxon signed rank test*, as the sign of the differences is also involved.

As for the sign test, the differences for all the pairs is calculated, then they are ordered and finally the positive differences $S_1, S_2, \ldots, S_p$ and the negative differences $R_1, R_2, \ldots, R_m$ $(p + m = n)$ are separated.

The statistic used to show whether both samples have the same position is defined as the sum of the S_i's:

$$V_S = \sum_{i=1}^{p} S_i$$

The expectation and the variance of V_s are:

$$E(V_S) = \frac{n(n+1)}{4} \quad \text{and} \quad V(V_S) = \frac{n(n+1)(2n+1)}{24}$$

Where they might be ties among the differences, or null differences for certain pairs, we have:

$$E(V_S) = \frac{n(n+1) - d_0(d_0+1)}{4}$$

$$V(V_S) = \frac{[n(n+1)(2n+1) - d_0(d_0+1)(2d_0+1)]}{24} - \frac{\sum_{i=1}^{nd}(d_i^3 - d_i)}{48}$$

where d_0 is the number of null differences, nd the number of distinct differences, and d_i the number of values corresponding to the i'th distinct difference value (it is the same as considering that the d_i's are the number of ties for the i'th distinct difference value).

Where there are no null differences or ties among the differences, if $n \leq 100$, XLSTAT calculates an exact p-value (Lehmann, 1975). Where there are ties, a normal approximation is used. We have:

$$P(V_S \leq \nu) \approx \Phi\left(\frac{\nu - E(V_S) + c}{\sqrt{V(V_S)}}\right)$$

where F is the distribution function for the standardized normal distribution, and c is a continuity correction used to increase the quality of the approximation (c is $\frac{1}{2}$ or $-\frac{1}{2}$ depending on the nature of the test). The approximation is more reliable the higher n is.

A Monte Carlo approximation of the p-value is also possible for this test.

Computation of the p-values

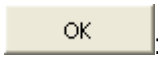
To compute the p-values corresponding to the various statistics, XLSTAT offers several alternatives:

- Asymptotic method: The p-value is obtained using the asymptotic approximation of the distribution of the statistic. The reliability of the approximation depends on the number of samples and on the number of measures per sample.
- Exact method: The computation of the p-value is based on the exact distribution of the statistic.
- Monte Carlo method: The computation of the p-value is based on random resamplings. The user must set the number of resamplings. A confidence interval on the p-value is provided. The more resamplings are performed, the better the estimation of the p-value.

In order to avoid freezing Excel because of too long computations, it is possible with the two latter methods to set the maximum time that should be spent computing the p-value.

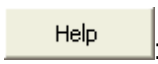
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.

: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data / Sample 1: If the format of the selected data is "one column per variable", select the data for the various samples in the Excel worksheet. If the format of the selected data is "one column per sample" or "paired samples", select a column of data corresponding to the first sample.

Sample identifiers / Sample 2: If the format of the selected data is "one column per variable", select the data identifying the two samples to which the selected data values correspond. If the format of the selected data is "one column per sample" or "paired samples", select a column of data corresponding to the second sample.

Data format: choose the data format.

- **One column/row per sample**: Activate this option to select one column (or row in row mode) per sample.
- **One column/row per variable**: Activate this option for XLSTAT to carry out as many tests as there are columns/rows, given that each column/row must contain the same number of rows/columns and that a sample identifier which enables each observation to be assigned to a sample must also be selected.
- **Paired samples**: Activate this option to carry out tests on paired samples. You must then select a column (or row in row mode) per sample, all the time ensuring that the samples are of the same size.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

Mann-Whitney test: Activate this option to run the Mann-Whitney test (see description).

Sign test: Activate this option to use sign test (see description).

Wilcoxon signed rank test: Activate this option to use Wilcoxon signed rank test (see description).

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Hypothesized difference (D): Enter the value of the supposed difference between the samples.

Significance level (%): Enter the significance level for the test (default value: 5%).

Depending on the test that is being used, several methods can be available to compute the **p-value**. Choose among the **asymptotic**, **exact** or **Monte Carlo** methods (see the description section for more information). In the case of the exact and Monte Carlo method you can set the maximum time you want to spend computing the p-value.

Continuity correction: Activate this option if you want XLSTAT to use the continuity correction when computing the asymptotic p-value (see description).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected samples.

Charts tab:

Dominance diagram: Activate this option to display a dominance diagram in order to make a visual comparison of the samples.

Results

The results displayed by XLSTAT relate to the various statistics of the tests selected and the interpretation arising from these.

Example

A tutorial showing how to use the Mann-Whitney test is available on XLSTAT Help Center:

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

A tutorial showing how to run a Sign test and a Wicoxon-rank-signed test is available on XLSTAT Help Center:

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

References

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.

Comparison of k samples (Kruskal-Wallis, Friedman, ...)

Use this tool to compare k independent samples (Kruskal-Wallis test and Dunn's procedure) or paired samples (Friedman's test and Nemenyi's procedure, GPU accelerated).

In this section:

Description

To get round the assumption that a sample is normally distributed required for using multiple comparison tests (offered in XLSTAT after an ANOVA), non- parametric tests were proposed.

As for parametric tests, a distinction is made between independent samples (for example a comparison of crop yields from fields with similar properties but treated with three different types of fertilizer), from cases where they are paired (for example if comparing the scores given by 10 judges to 3 different products).

Comparison of k independent samples

The **Kruskal-Wallis test** is often used as an alternative to the ANOVA where the assumption of normality is not acceptable. It is used to test if k samples ($k \geq 2$) come from the same population or populations with identical properties as regards a position parameter (the position parameter is conceptually close to the median, but the Kruskal-Wallis test takes into account more information than just the position given by the median).

If M_i the position parameter for sample i, the null H0 and alternative Ha hypotheses for the Kruskal-Wallis test are as follows:

- H0: $M_1 = M_2 = \ldots = M_k$
- Ha: There is at least one pair (i, j) such that $M_i \neq M_j$

The calculation of the K statistic from the Kruskal-Wallis test involves, as for the Mann-Whitney test, the rank of the observations once the k samples (or groups) have been mixed. K is defined by:

$$K = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

where n_i is the size of sample i, N is the sum of the n_i's, and R_i is the sum of the ranks for sample i.

When $k = 2$, the Kruskal-Wallis test is equivalent to the Mann-Whitney test and K is equivalent to W.

When there are ties, the mean ranks are used for the corresponding observations as in the case of the Mann-Whitney test. K is then given by:

$$K = \frac{\frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)}{1 - \sum_{i=1}^{nd} (d_i^3 - d_i)/(N^3 - N)}$$

where nd is the number of distinct values and d_i the number of observations for each of the values.

The distribution of the K statistic can be approximated by a Chi-square distribution with $(k-1)$ degrees of freedom. This approximation is reliable, except when N is small. The p-values associated with K, which for the exact case depends on the statistic K and the k sizes of the samples, have been tabulated for the case where $k = 3$ (Lehmann 1975, Hollander and Wolfe 1999).

Comparison of k paired samples

The **Friedman test** is a non-parametric alternative to the two-way ANOVA where the assumption of normality is not acceptable. It is used to test if k paired samples ($k \geq 2$) of size n, come from the same population or from populations having identical properties as regards the position parameter. As the context is often that of the two-way ANOVA factors, we sometimes speak of the Friedman test with k treatments and n blocks.

If M_i is the position parameter for sample i, the null H0 and alternative Ha hypotheses for the Friedman test are as follows:

- H0: $M_1 = M_2 = \ldots = M_k$
- Ha: There is at least one pair (i, j) such that $M_i \neq M_j$

Let n be the size of k paired samples. The Q statistic from the Friedman test is given by:

$$Q = \frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3n(k+1)$$

where R_i is the sum of the ranks for sample i.

Where there are ties, the average ranks are used for the corresponding observations. Q is then given by:

$$Q = \frac{\frac{12}{nk(k+1)}\sum_{i=1}^{k} R_i^2 - 3n(k+1)}{1 - \sum_{j=1}^{n}\sum_{i=1}^{nd(j)}(d_{ij}^3 - d_{ij})/n/(k^3-k)}$$

where $nd(j)$ is the number of distinct values for block j, and d_{ij} the number of observations for each of the values.

As for the Kruskal-Wallis test, the p-value associated with a given value of Q can be approximated by a Chi-square distribution with $(k-1)$ degrees of freedom. This approximation is reliable when $k \times n$ is greater than 30, the quality also depending on the number of ties. The p-values associated with Q have been tabulated for (k = 3, n = 15) and (k = 4, n = 8) (Lehmann 1975, Hollander and Wolfe 1999).

Computation of the p-values

To compute the p-values corresponding to the various statistics, XLSTAT offers several alternatives:

- Asymptotic method: The p-value is obtained using the asymptotic approximation of the distribution of the statistic. The reliability of the approximation depends on the number of samples and on the number of measures per sample.
- Exact method: The computation of the p-value is based on the exact distribution of the statistic. This method is available for the Kruskal-Wallis test when there are no ties.
- Monte Carlo method: The computation of the p-value is based on random resamplings. The user must set the number of resamplings. A confidence interval on the p-value is provided. The more resamplings are performed, the better the estimation of the p-value.

In order to avoid freezing Excel because of too long computations, it is possible with the two latter methods to set the maximum time that should be spent computing the p-value.

Multiple pairwise comparisons

Whether for the Kruskal-Wallis or the Friedman test, if the p-value is such that the H0 hypothesis has to be rejected, then at least one sample (or group) is different from another. To identify which samples are responsible for rejecting H0, **multiple comparison** procedures can be used.

For the Kruskal-Wallis test, three multiple comparison methods are available:

- Dunn (1963): the method based on the comparison of the mean of the ranks of each treatment, the ranks being those used for the computation of K. The normal distribution is used as the asymptotic distribution of the standardized difference of the mean of the ranks.

- Conover and Iman (1999): close to Dunn's method, this method uses a Student distribution. It corresponds to a t test performed on the ranks.
- Steel-Dwass-Critchlow-Fligner (1984): This more complex method is recommended by Hollander (1999). It requires the recalculation of the ranks for each combination of treatments. The Wij statistic is calculated for each combination. XLSTAT then calculates the corresponding p-value using the asymptotic distribution.

For the Friedman test also, three multiple comparison methods are proposed:

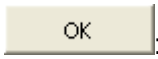
- Nemenyi (1963): this method is close to that of Dunn, but takes into account data matching.
- Conover and Iman (1999): close to Dunn's method, this method uses a Student distribution. It corresponds to a t test performed on the ranks.
- Siegel and Castellan (1988) (or Bonferroni - Dunn): this method based on the comparison of the mean of the ranks of each treatment, the ranks being those used for the computation of K. The normal distribution is used as the asymptotic distribution of the standardized difference of the mean of the ranks.

For the methods of Dunn, Conover and Iman, and Siegel and Castellan, to take into account the fact that there are $k(k-1)/2$ possible comparisons, the correction of the significance level proposed by Bonferroni can be applied. The significance level used for pairwise comparisons is:

$$\alpha' = \frac{2\alpha}{k(k-1)}$$

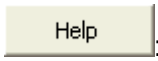
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.

: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the

arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data: If the format of the selected data is "one column per variable", select the data for the various samples in the Excel worksheet. If the format of the selected data is "one column per sample" or "paired samples", select the columns of data corresponding to the various samples.

Sample identifiers: If the format of the selected data is "one column per variable", select the data identifying the k samples to which the selected data values correspond.

Data format: choose the data format.

- **One column/row per sample**: Activate this option to select one column (or row in row mode) per sample.
- **One column/row per variable**: Activate this option for XLSTAT to carry out as many tests as there are columns/rows, given that each column/row must contain the same number of rows/columns and that a sample identifier which enables each observation to be assigned to a sample must also be selected.
- **Paired samples**: Activate this option to carry out tests on paired samples. You must then select a column (or row in row mode) per sample, all the time ensuring that the samples are of the same size.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

Kruskal-Wallis test: Activate this option to run the Kruskal-Wallis test (see description).

Friedman test: Activate this option to run a Friedman test (see description).

Multiple pairwise comparisons: Activate this option to compute multiple pairwise comparisons (see description).

- **Bonferronicorrection**: Activate this option to use the Bonferroni corrected significance level for the multiple comparisons.

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Significance level (%): Enter the significance level for the test (default value: 5%).

Depending on the test that is being used, several methods can be available to compute the p-value. Choose among the **asymptotic**, **exact** or **Monte Carlo** methods (see the description section for more information). In the case of the exact and Monte Carlo method you can set the maximum time you want to spend computing the p-value.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected samples.

Results

The results displayed by XLSTAT relate to the various statistics of the tests selected and the interpretation arising from these. Results of multiple comparisons are then displayed to identify the treatments responsible for rejecting the null hypothesis, if it has been rejected.

Example

A tutorial showing how to use the Kruskal-Wallis test is available on XLSTAT Help Center:

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

A tutorial showing how to use the Friedman's test is available on XLSTAT Help Center:

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

References

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.

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.

Durbin-Skillings-Mack test

Use this tool to test if k treatments being measured within a (balanced or not) incomplete block design are identical or different (GPU accelerated).

In this section:

Description

The goal of the test proposed by Durbin (1951) is to allow analyzing rigorously the results of a study carried out within the framework of a balanced incomplete block design (BIBD), using a nonparametric procedure – thus not making any assumption on the distribution of the measurements. Skillings and Mack (1981) suggested an extension of this approach for more general incomplete block designs.

Block designs

A block design is a design in which we study the influence of two factors on one or more phenomena. We know that one factor has an impact that we cannot control, but that is not of interest. So we want to ensure that this factor does not disturb the analysis that we perform once the data have been collected. For this we make sure that the various levels of other factors are well represented in each block.

The blocking factor can correspond to judges evaluating products, and the factor of interest would then be the products being studied.

A complete block design is a design in which all levels of the factors of interest are present once within each block. For a sensory design, this corresponds to a design where all products are seen once by each judge.

In an incomplete block design, all levels of the factors of interest are not present for all levels of the blocking factor. It is balanced if each level of the factor of interest is present a same number of times r in the design, and if each pair of levels of each factor is present the same number of times λ .

If t is the number of treatments, b the number of blocks, k the number of treatments measured within each block, we show that the following conditions are necessary (but not sufficient) to

have a balanced incomplete block design:

$$bk = tr$$

and

$$r(k-1) = \lambda(t-1)$$

The Durbin and Skillings-Mack tests

The Durbin and Skillings-Marck tests are an extension of the Friedman test (1937) that can only be used in the case of complete block designs.

If $T_1, T_2, \ldots, T_t$ correspond to the t treatments, as for the Friedman test, the null and alternative hypotheses used in the test are:

- H_0 : The t treatments are not different.
- H_a : At least one of the treatments is different from another.

The Durbin statistic is given by

$$Q = \frac{12(t-1)}{rt(k-1)(k+1)} \sum_{j=1}^{t} \left(R_j - \frac{r(k+1)}{2} \right)^2$$

where R_j is the sum over the b blocks for treatment j on ranks R_{ij} of block i.

In the case where there are ties in one or more blocks, the variance must be corrected. We then have:

$$Q = \frac{(t-1)}{A-C} \left(\left(\sum_{j=1}^{t} R_j^2 \right) - rC \right) \text{ with } A = \sum i = 1^b \sum_{j=1}^{t} R_{ij}^2 \text{ and } C = \frac{bk(k+1)}{4}$$

This statistic has the property to be asymptotically distributed according to a χ^2 distribution with $t-1$ degrees of freedom. Alvo and Cabilio (1995) propose a modified statistic with, according to Conover (1999), better asymptotic properties

$$F = \frac{\frac{Q}{(t-1)}}{(b(k-1)-Q)(b(k-1)-t+1)}$$

This statistic has the property to be asymptotically distributed according to a Fisher F with $t-1$ and $b(k-1)-t+1$ degrees of freedom.

The computation of Skillings and Mack T statistic which allows to treat unbalanced incomplete block designs is more complex. The missing values are replaced by an average of the ranks,

and a compensatory weight is applied to blocks with missing values. The T statistic asymptotically follows a χ^2 distribution with $t-1$ degree of freedom.

Computation of the p-values

To compute the p-values corresponding to the various statistics, XLSTAT offers several alternatives:

- Asymptotic method: The p-value is obtained using the asymptotic approximation of the distribution of the Q and F statistics. The reliability of the approximation depends on the number of treatments and on the number of blocks.
- Monte Carlo method: The computation of the p-value is based on random resamplings. The user must set the number of resamplings. A confidence interval on the p-value is provided. The more resamplings are performed, the better the estimation of the p-value.

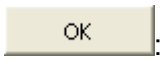
In order to avoid freezing Excel because of too long computations, it is possible with the two latter methods to set the maximum time that should be spent computing the p-value.

Multiple pairwise comparisons

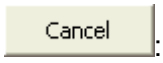
If the p-value is such that the H_0 hypothesis has to be rejected, then at least one treatment is different from another. To identify which treatments are responsible for rejecting H_0, a **multiple comparison** procedure can be used. XLSTAT allows to use for the Durbin test, which is the procedure suggested by Conover (1999). In the case of non balanced incomplete block designs, Conover's procedure is also used with Bonferroni's correction.

Dialog box

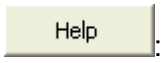
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Subjects/Treatments table: Select a table where each row (or column if in column mode) corresponds to a block, and each column (or row in row mode) corresponds to a treatment. If headers have been selected with the data, make sure the "Treatment labels" option is checked.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Treatment labels: Activate this option if headers have been selected with the input data.

Multiple pairwise comparisons: Activate this option to compute multiple pairwise comparisons.

Options tab:

Significance level (%): Enter the significance level for the test (default value: 5%).

To compute the p-value, you can either choose the asymptotical approximation method or the Monte Carlo resamplings based method (see the description section). In the latter case, you can set the number of resamplings you want to make, and the maximum time you want XLSTAT to spend on making the resamplings.

Outputs tab:

Descriptive statistics: Activate this option to compute and display the statistics that correspond to each treatment.

Results

Descriptive statistics: This table displays the descriptive statistics that correspond to the k treatments.

The results that correspond to the Durbin test (in the case of balanced incomplete block design) or to the Skillings-Mack test (in the case of incomplete block designs) are then displayed,

followed by a short interpretation of the test. Results of multiple comparisons are then displayed to identify the treatments responsible for rejecting the null hypothesis, if it has been rejected.

Example

An example showing how to run a Durbin test is available on XLSTAT Help Center:

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

References

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.

Page test

Use this tool to test if k treatments being measured within a (balanced or not) incomplete block design are identical or if a sorting of the treatments is possible (GPU accelerated).

In this section:

Description

The goal of the test proposed by Page (1963) is to allow analyzing rigorously the results of a study carried out within the framework of a complete design, to verify if a series of several treatments should be considered as not different, or if alternatively a ranking of the treatment makes sense. The Page test is a nonparametric method, thus not making any assumption on the distribution of the measurements. This test differs from the Friedman test by the fact that the alternative hypothesis is a ranking of the treatments and not only a difference. This test has been extented to the case of incomplete blocks by Alvo and Cabilio (2005).

Block designs

A block design is a design in which we study the influence of two factors on one or more phenomena. We know that one factor has an impact that we cannot control, but that is not of interest. So we want to ensure that this factor does not disturb the analysis that we perform once the data are collected. For this we make sure that the various levels of other factors are well represented in each block.

The blocking factor can correspond to judges evaluating products, and the factor of interest would then be the products being studied.

A complete block design is a design in which all levels of the factors of interest are present once within each block. For a sensory design, this corresponds to a design where all products are seen once by each judge.

In an incomplete block design, all levels of the factors of interest are not present for all levels of the blocking factor. It is balanced if each level of the factor of interest is present a same number of times r in the design, and if each pair of levels of each factor is present the same number of times l.

If t is the number of treatments, b the number of blocks, k the number of treatments measured within each block, we show that the following conditions are necessary (but not sufficient) to have a balanced incomplete block design:

$$bk = tr$$

and

$$r(k-1) = l(t-1)$$

The Page test

If $T_1, T_2, \ldots, T_t$ correspond to the t treatments, the null and alternative hypotheses used in the test are:

- H0 : The t treatments are not significantly different.
- Ha : $T_1 \geq T_2 \geq \ldots \geq T_t$

or

- Ha : $T_1 \leq T_2 \leq \ldots \leq T_t$

Where, for the alternative hypotheses, at least one inequality is strict.

The statistic suggested by Page is given by:

$$L = \sum_{j=1}^{t}(jR_j)$$

Page tabulated that statistic, and also gave an asymptotic approximation with a Chi-square distribution with 1 degree of freedom. Conover (1999) uses the following statistic that follows a standard normal distribution:

$$z = \frac{12L - 3bt(t+1)^2}{\sqrt{bt^2(t^2-1)(t+1)}}$$

Where b is the number of blocs and t the number of treatments. In case there are ties for a given block, the variance term is adapted.

In the case of incomplete blocks, Alvo and Cabilio (2005) suggest an alternative statistic which value is identical in the case of complet blocks and that has the same asymptotical properties.

Computation of the p-values

To compute the p-values corresponding to the various statistics, XLSTAT offers several alternatives:

- Asymptotic method: The p-value is obtained using the asymptotic approximation of the distribution of the *z* statistics. The reliability of the approximation depends on the number of treatments and on the number of blocks.
- Monte Carlo method: The computation of the p-value is based on random resamplings. The user must set the number of resamplings. A confidence interval on the p-value is provided. The more resamplings are performed, the better the estimation of the p-value.

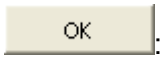
In order to avoid freezing Excel because of too long computations, it is possible with the two latter methods to set the maximum time that should be spent computing the p-value.

Multiple pairwise comparisons

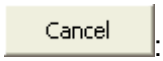
If the p-value is such that the H0 hypothesis has to be rejected, then at least one treatment is different from another. To identify which treatment(s) is/are responsible for rejecting H0, a **multiple comparison** procedure can be used, XLSTAT allows using the procedure suggested by Cabilio and Peng (2008), with two alternative ways to compute the p-value of the paired comparisons. It can either use the normal approximation of a Monte Carlo based -pvalue.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Subjects/Treatments table: Select a table where each row (or column if in column mode) corresponds to a block, and each column (or row in row mode) corresponds to a treatment. If

headers have been selected with the data, make sure the "Treatment labels" option is checked.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Treatment labels: Activate this option if headers have been selected with the input data.

Multiple pairwise comparisons: Activate this option to compute multiple pairwise comparisons. Two methods are proposed: Cabilio and Peng using an asymptotical p-value or the same procedure with p-values computed through Monte Carlo resamplings (see the description section for more information).

Options tab:

Significance level (%): Enter the significance level for the test (default value: 5%).

To compute the p-value, you can either choose the asymptotical approximation method or the Monte Carlo resamplings based method (see the description section). In the latter case, you can set the number of resamplings you want to make, and the maximum time you want XLSTAT to spend on making the resamplings.

Outputs tab:

Descriptive statistics: Activate this option to compute and display the statistics that correspond to each treatment.

Results

Descriptive statistics: This table displays the descriptive statistics that correspond to the k treatments.

The results that correspond to the Page test (complete block design) or to the Alvo and Cabilio variant of the test (incomplete block designs) are then displayed, followed by a short interpretation of the test. Results of multiple comparisons are then displayed to identify the treatments responsible for rejecting the null hypothesis, if it has been rejected.

Example

An example showing how to run a Page test is available on XLSTAT Help Center:

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

References

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.

Cochran's Q test

Use this tool to compare $k \geq 2$ paired samples which values are binary (GPU accelerated).

In this section:

Description

The Cochran's Q test is presented using two different approaches. Some authors present it as a particular case of the Friedman's test (comparison a k paired samples) when the variable is binary (Lehmann, 1975), while other present it as a marginal homogeneity test for a k-dimensional contingency table (Agresti, 1990).

As a consequence, the null H0 and alternative Ha hypotheses for the Cochran's Q test can either be,

- H0: The k treatments are not different.
- Ha: At least on of the treatment is different from another.

or,

- H0: the k distributions are marginally homogeneous.
- Ha: the k distributions are marginally inhomogeneous.

XLSTAT uses the first approach, as it is the most used. The term "treatment" has been chosen for the k samples that are being compared.

Two possible formats are available for the input data:

- You can select data in a "raw" format. In this case, each column corresponds to a treatment and each row to a subject (or individual, or bloc).
- You can also select the data in a "grouped" format. Here, each column corresponds to a treatment, and each row corresponds to a unique combine of the k treatments. You then need to select the frequencies corresponding to each combine (field "Frequencies" in the dialog box).

Computation of the p-values

To compute the p-values corresponding to the various statistics, XLSTAT offers several alternatives:

- Asymptotic method: The p-value is obtained using the asymptotic approximation of the distribution of the statistic. The reliability of the approximation depends on the number of samples and on the number of measures per sample.
- Exact method: The computation of the p-value is based on the exact distribution of the statistic.
- Monte Carlo method: The computation of the p-value is based on random resamplings. The user must set the number of resamplings. A confidence interval on the p-value is provided. The more resamplings are performed, the better the estimation of the p-value.

In order to avoid freezing Excel because of too long computations, it is possible with the two latter methods to set the maximum time that should be spent computing the p-value.

Multiple pairwise comparisons

If the p-value is such that the H0 hypothesis has to be rejected, then at least one treatment is different from another. To identify which samples are responsible for rejecting H0, **multiple comparison** procedures can be used. XLSTAT suggests using the Marascuilo and McSweeney (1977) option, while a McNemar test with the Bonferroni correction is also available.

Multiple pairwise comparisons

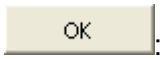
When the H0 hypothesis of the Cochran test is rejected, it is concluded that at least one treatment significantly differs from the others. It may be interesting to compare treatments pairwise. For this XLSTAT proposes two different methods:

- Method **Critical difference (Sheskin)** explained in Sheskin (2011) and initially developed by Marascuilo and McSweeney (1977). This method calculates a critical value CD. When the difference in proportion between two treatments is greater than CD, it is concluded that there is a significant difference between the two treatments.
- The **McNemar(Bonferroni)** procedure performs McNemar tests between the different treatment pairs and applies the Bonferroni correction. This correction consists of dividing the $alpha$ significance level by the total number of comparisons ($k \times (k-1)/2$). The p-values of McNemar tests are obtained by an asymptotic approximation of the distribution of calculated statistics. A continuity correction is also applied. Refer to the McNemar test section for more details.

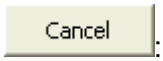
The **Critical difference (Sheskin)** method is the preferred option because the critical value is calculated taking into account all treatments as opposed to multiple McNemar tests which use only the data of the two treatments to be compared.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Subjects/Treatments table: Select a table where each row (or column if in column mode) corresponds to a subject, and each column (or row in row mode) corresponds to a treatment. If headers have been selected with the data, make sure the "Treatment labels" or "Labels included" is checked.

Data format:

- Subjects/Treatments table:
- **Raw**: Choose that option if the input data are in a raw format (as opposed to grouped).
- **Grouped**: Choose that option if your data correspond to a summary table where each row corresponds to a unique combine of treatments. You then need to select the frequencies that correspond to each combine (see "Frequencies" below).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Treatment labels: Activate this option if headers have been selected with the input data.

Weights: Select the weights that correspond to the combines of treatments. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Treatment labels" option is activated.

Multiple pairwise comparisons: Activate this option to compute multiple pairwise comparisons. Two methods are proposed: **Critical difference (Sheskin)** or **McNemar (Bonferroni)** (see the description section for further details).

Options tab:

Significance level (%): Enter the significance level for the test (default value: 5%).

Depending on the test that is being used, several methods can be available to compute the p-value. Choose among the asymptotic, exact or Monte Carlo methods (see the description section for more information). In the case of the exact and Monte Carlo method you can set the maximum time you want to spend computing the p-value.

Outputs tab:

Descriptive statistics: Activate this option to compute and display the statistics that correspond to each treatment.

Results

Descriptive statistics: This table displays the descriptive statistics that correspond to the k treatments.

The results that correspond to the Cochran's Q test are then displayed, followed by a short interpretation of the test.

Example

An example showing how to run a Cochran's Q test is available on XLSTAT Help Center:

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

References

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.

McNemar’s test

Use this tool to compare 2 paired samples which values are binary. The data can be summarized in a 2x2 contingency table.

In this section:

Description

The McNemar’s test is a special case of the Cochran’s Q test when there are only two treatments. As for the Cochran’s Q test, the variable of interest is binary. However, the McNemar’s test has two advantages:

- Obtaining an exact p-value is possible (Lehmann, 1975);

- The data can be summarized in a 2x2 contingency table.

In the case of the two-tailed (or two-sided) test, the null (H0) and alternative (Ha) hypotheses are:

- H0: Treatment 1 = Treatment 2
- Ha: Treatment 1 $\neq$ Treatment 2

In the one-tailed case, you need to distinguish the left-tailed (or lower- tailed or lower one-sided) test and the right-tailed (or upper-tailed or upper one-sided) test. In the left-tailed test, the following hypotheses are used:

- H0: Treatment 1 = Treatment 2
- Ha: Treatment 1 < Treatment 2

In the right-tailed test, the following hypotheses are used:

- H0: Treatment 1 = Treatment 2
- Ha: Treatment 1 > Treatment 2

Three possible formats are available for the input data:

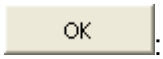
- You can select data in a "raw" format. In this case, each column corresponds to a treatment and each row to a subject (or individual, or bloc).

- You can also select the data in a "grouped" format. Here, each column corresponds to a treatment, and each row corresponds to a unique combine of the k treatments. You then need to select the frequencies corresponding to each combine (field "Frequencies" in the dialog box).

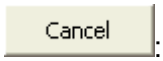
- You can also select a contingency table with two rows and two columns. In the case where you choose this, the first and second treatments are respectively considered as corresponding to the rows and the columns. The positive response cases (or successes) are considered as corresponding to the first row of the contingency table for the first treatment, and to the first column for the second treatment.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Subjects/Treatments table / Contingency table (2x2): In the case of a "Subjects/Treatments table), select a table where each row (or column if in column mode) corresponds to a subject, and each column (or row in row mode) corresponds to a treatment. In the case of a "Contingency table", select the contingency table. If headers have been selected with the data, make sure the "Treatment labels" or "Labels included" is checked.

Data format:

- **Subjects/Treatments table**: Choose this option if the data correspond to a Subjects/Treatments table.
- **Raw**: Choose that option if the input data are in a raw format (as opposed to grouped).
- **Grouped**: Choose that option if your data correspond to a summary table where each row corresponds to a unique combine of treatments. You then need to select the frequencies that correspond to each combine (see "Frequencies" below).
- **Contingency table (2x2)**: Activate this option if your data are available in a 2x2 contingency table.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Treatment labels/Labels included: Activate this option if headers have been selected with the input data. In the case of a contingency table, the row and column labels must be selected if this option is checked.

Weights: Select the weights that correspond to the combines of treatments. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Treatment labels" option is activated.

Positive response code: Enter the value that corresponds to a positive response in your experiment.

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Significance level (%): Enter the significance level for the test (default value: 5%).

Exact p-value: Activate this option to compute the exact p-value.

Outputs tab:

This tab is only visible if the "Subjects/Treatments table" format has been chosen.

Descriptive statistics: Activate this option to compute and display the statistics that correspond to each treatment.

Contingency table: Activate this option to display the 2x2 contingency table.

Results

Descriptive statistics: This table displays the descriptive statistics that correspond to the two treatments.

Contingency table: The 2x2 contingency table built from the input data is displayed.

The results that correspond to the McNemar's test are then displayed, followed by a short interpretation of the test.

Example

An example showing how to run a McNemar's test is available on XLSTAT Help Center:

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

References

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.

Cochran-Mantel-Haenszel Test

Use this tool to test the hypothesis of independence on a series of contingency tables corresponding to an experiment crossing two categorical variables, with a control variable taking multiple values.

In this section:

Description

Imagine the case of a laboratory working on a new antifungal agent. In order to define the appropriate dose and dosage form, an experiment is conducted with four dose levels and two different dosage forms (ointment or shower gel). For each dose level, the test is performed on about twenty patients, divided equally for each presentation. The experimenters record for each patient whether the treatment is effective or not. The results are thus in the form of a contingency table with three dimensions, or more simply in the form of 4 two-way contingency tables. The variable corresponding to the dose is the control variable.

One could be willing to do a test of independence on the table resulting from the sum of 4 contingency tables, however in this case one could conclude that there is independence for the sole reason that the sub-contingency table with the largest number of respondents corresponds a case of independence, while the other tables do not at all.

Cochran (1954) then Mantel and Haenszel (1959) developed a test that allows to test whether there is independence or not between the rows and columns of the contingency tables, taking into account the fact that the tables are independent of each other (for each dose the patients are different), and by conditioning on the marginal sums of each table, as in the standard test of independence on contingency tables.

The test commonly named the Cochran-Mantel-Haenszel (CMH) test is based on the M^2 statistic defined by:

$$M^2 = \frac{\left(\left|\sum_{i=1}^{k}\left(n_{11i} - n_{1+i}n_{+1i}/n_{++i}\right)\right| - \frac{1}{2}\right)^2}{\sum_{i=1}^{k} n_{1+i}n_{2+i}n_{+1i}n_{+2i}/\left(n_{++i}^2\left(n_{++i}^2 - 1\right)\right)}$$

This statistic follows asymptotically a chi-square distribution with 1 degree of freedom. Knowing M^2, we can therefore compute the p-value, and knowing the risk of Type I, α, we can

determine the critical value. It is also possible, as for the test of independence on a contingency table, to calculate the exact p-value, if the contingency tables are of size 2x2. The use of absolute value and the subtraction of $-\frac{1}{2}$ and the division by $(n_{++i}^2 - 1)$ instead of n_{++i}^2 corresponds to a continuity correction proposed by Mantel and Haenszel. Its use is strongly recommended. With XLSTAT you have the choice to use it (default) or not.

It may be noted that the numerator measures for the upper left cell the difference between the actual value and the expected value corresponding to independence, and that then sum these differences. If the differences are in opposite directions from one table to another we could therefore conclude that there is independence while there is dependence in each table (Type II error). This situation happens when there is a three-way interaction between the three variables. This test is to be used with caution.

The Cochran-Mantel-Haenszel test has been generalized by Birch (1965), Landis *et al.* (1978) and Mantel and Byar (1978) to the case of RxC contingency tables where R and C can be greater than 2. The computation of M^2 is more complex, but it still leads to a statistic that asymptotically follows a χ^2 with $(L-1)(C-1)$ degrees of freedom.

It is recommended to perform separately from the CMH test, the analysis of the Cramer's V for the individual contingency tables to get an idea of their contribution to independence. XLSTAT displays automatically for each contingency table, a table with the Cramer's V, the χ^2 and the corresponding p-values (exact for 2x2 tables and asymptotic for higher dimensional tables) where possible, that is, when there are no null marginal sums.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

  : Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Contingenc y tables: If the data format selected is "contingency tables", select the k contingency tables, and then specify the value of k by entering the **number of strata**.

Variable 1: If the data format selected is "variables", select the data corresponding to the first qualitative variable used to construct contingency tables.

Variable 2: If the data format selected is "variables", select the data corresponding to the second qualitative variable used to construct contingency tables.

Strata: If the selected data format is "variables", select the data corresponding to the various strata.

Data format: Select the data format.

- **Contingency tables**: Activate this option if your data are available as a set of k contingency tables one under the other.
- **Variables**: Activate this option if your data are available as two qualitative variables with one row for each observation and one variable corresponding to the various strata (control variable).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

Options tab:

Significance level (%): Enter the significance level for the test (default value: 5%).

Exact p-values: Activate this option to compute the exact p-values when possible (see description).

Alternative hypothesis: Choose the alternative hypothesis to be used for the test in the case of an exact p-value computed on a set of 2x2 tables (see description).

Common odds ratio: Enter the value of the assumed common odds-ratio.

Continuity correction: Activate this option if you want XLSTAT to use the continuity correction if the exact p-values calculation has not been requested or is not possible (see description).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Results

The results that correspond to the Cochran-Mantel-Haenszel test are displayed, followed by a short interpretation of the test.

Example

A tutorial showing how to use the Cochran-Mantel-Haenszel test is available on XLSTAT Help Center:

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

References

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.

One-sample runs test

Use this tool to test whether a series of binary events is randomly distributed or not.

In this section:

Description

The first version of this nonparametric test was presented by Mood (1940) and is based on the same runs statistic as the two-sample test by Wald and Wolfowitz (1940), which is why this test is sometimes mistakenly referred to as the Wald and Wolfowitz runs test. However, the article by Mood makes reference to the article by Wald and Wolfowitz and the asymptotic distribution of the statistic uses also the results given by these authors.

A run is a sequence of identical events, preceded and succeeded by different or no events. The runs test used here applies to binomial variables only. For example, in ABBABBB, we have 4 runs (A, BB, A, BBB).

XLSTAT accepts as input, continuous data or binary categorical data. For continuous data, a cut-point must be chosen by the user so that the data are transformed into a binary sample.

A sample will be considered as randomly distributed if no particular structure can be identified. Extreme cases are repulsion, where you have all observations of one kind on the left, and all the remaining observations on the right, and alternation where the elements of the two kinds are alternating as much as possible. With the previous case, repulsion would give "AABBBBB" or "BBBBBAA", and alternation "BABABBB" or "BABBABB" or "BBABABB" or "BBABBAB" or "BBBABAB".

In the case of the two-tailed (or two-sided) test, the null (H0) and alternative (Ha) hypotheses are:

- H0: data are randomly distributed.
- Ha: data are not randomly distributed.

In the one-tailed case, you need to distinguish the left-tailed (or lower- tailed or lower one-sided) test and the right-tailed (or upper-tailed or upper one-sided) test. In the left-tailed test, the following hypotheses are used:

- H0: data are randomly distributed.
- Ha: there is repulsion between the two types of events.

In the right-tailed test, the following hypotheses are used:

- H0: data are randomly distributed.
- Ha: The two types of events are alternating.

The expectation of the number of runs R is given by:

E(R) = 2mn/N

where m is the number of events of type 1, and n the number of events of type 2, and N is the total sample size.

The variance of the number of runs R is given by:

V(R) = 2mn(2mn – N)/[N²(N-1)]

The minimum value of R is always 2. The maximum value is given by 2Min(m, n) - t, where t is 1 if m=n, and 0 if not.

If r is the number of runs measured on the sample, it was shown by Wald and Wolfowitz that asymptotically, when m or n tend to infinity,

$$\frac{(r - E(R))}{\sqrt{V(R)}} \rightarrow N(0,1)$$

where $N(0,1)$ is the standard normal distribution.

XLSTAT offers three ways to compute the p-values. You can compute the p-value based on:

- The exact distribution of R,
- The asymptotic distribution of R,
- An approximated distribution based on P Monte Carlo permutations. As the number of possible permutations is high (it is equal to N!), P must be set to a high value so that the approximation is fine.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data: Select a column (or row in row mode) of data corresponding to the series of data to analyze.

Data type: Select the data type.

- **Quantitative**: Activate this option to select one column (or row in row mode) of quantitative data. The data will then be transformed on the basis of the cut point (see below).
- **Qualitative**: Activate this option to select one column (or row in row mode) of binary data.

Cut point: Choose the type of value that will be used to discretize the continuous data into a binary sample.

- **Mean**: Observations are split into two groups depending on whether there are lower or greater than the mean.
- **Median**: Observations are split into two groups depending on whether there are lower or greater than the median.
- **User defined**: Select this option to enter the value used to transform the data and enter that value. The observations are split into two groups depending on whether there are lower or greater than the given value.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Significance level (%): Enter the significance level for the test (default value: 5%).

Exact p-value: Activate this option if you want XLSTAT to calculate the exact p-value (see description).

Asymptotic p-value: Activate this option if you want XLSTAT to calculate the p-value based on the asymptotic approximation (see description).

- **Continuity correction**: Activate this option if you want XLSTAT to use the continuity correction when computing the asymptotic p-value.

Monte Carlo method: Activate this option if you want XLSTAT to calculate the p-value based on Monte Carlo permutations, and select the number of random permutations to perform.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Results

The results that correspond to the one-sample runs test are displayed, followed by a short interpretation of the test.

References

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.

Friedman-Rafsky test

Use this tool to compare the distributions of two samples with quantitative data.

In this section:

Description

The Friedman-Rafsky test is a nonparametric two-sample test. The null hypothesis is: $H0$: The X and Y samples follow the same distribution function i.e., $Fx = Fy$.

This test is a multivariate generalization of the Wald-Wolfowitz test. The test statistic if $p = 1$ (where p is the number of variables) is computed as following:

- join the two samples X of size m and Y of size n,
- sort the two samples in ascending order,
- replace each number by X or Y according to the sample it comes from,
- count r the number of successive X and Y,
- finally, compute:

$$ \frac { (r - E (R)) } { \sqrt { V (R) } } \rightarrow N (0,1), $$

where $E[R] = \frac{2mn}{N} + 1, Var(R) = \frac{2mn(2mn-m-n)}{N^2(N-1)}$ and $N = m + n$.

We reject $H0$ for small values of r.

For the multivariate case, "sorting" data is less intuitive. To counter this, Friedman and Rafsky use a minimum spanning tree to sort the data. The minimum spanning tree is a noncyclic graph that connects all the nodes together with the minimum total edge weight.

The test proceeds as follows:

- join the two samples X of size m and Y of size n,
- compute the distance matrix,
- compute the minimum spanning tree using the distance matrix as a complete graph,
- count r the number of edges that connect two nodes that do not come from the same sample to which we add 1.

The r number allows us to compute the test statistic as following: $$ W = \frac{r - E[R]} {Var(R)^{\frac{1}{2}}} \rightarrow N (0,1), $$

where $N(0,1)$ is the standard normal distribution and: $$ E[R] = \frac{2mn}{N}+1, $$

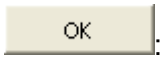
$$Var(R) = \frac{2mn}{N(N-1)}\{\frac{2mn-N}{N} + \frac{C-N+2}{(N-2)(N-3)}[N(N-1)-4mn+2]\},$$

where C is the number of edge pairs that share a common node.

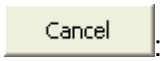
Again, we reject $H0$ for small values of r.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options, ranging from data selection to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Data format:

- **Separated samples**: Select two tables (a sample with m lines and another one with n lines) with possibly a different number of lines but with the same number p of columns.
- **Merged samples**: Select a table with $m+n$ lines and p quantitative variables. Select the (binary) data identifying the samples to which the selected data values correspond.

Distance: This option allows you to select the metric you wish to apply:

- **Euclidean distance**,
- **Manhattan distance**,
- **Chebychev distance**,
- **Canberra distance**.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet in the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selection (Observations/variables table, row labels, row weights, column weights) contains a label. Where the selection is a proximity matrix, if this option is activated, the first column must also include the object labels.

Minimum spanning tree / Algorithm: You can choose how to compute the minimum spanning tree between three methods:

- **Chazelle (Soft-Heap)** (by default): The Chazelle algorithm is a deterministic algorithm and is the fastest one (lowest asymptotic bounds) to compute a minimum spanning tree with a running time of $O(m\alpha(m,n))$ where α is the classical functional inverse of Ackermann's function.
- **Kruskal**: The Kruskal algorithm is one of the most used algorithms to compute a minimum spanning tree. This is a greedy algorithm that is to be preferred for small samples.
- **Boruvka**: This is the first algorithm invented to compute a minimum spanning tree. It is also a greedy algorithm.

Significance level (%): Enter the significance level for the tests (default value: 5%).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue the calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean**: Activate this option to estimate missing data by using the mean of the corresponding variables.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Minimum spanning tree: Activate this option to display on each line, a branch from the tree, the two nodes that the edge relies on and if the nodes come from the same sample.

Distance matrix: Activate this option to display the distance matrix.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. The number of observations per variable and per sample, the minimum, the maximum, the quartiles, the mean, the variance and the standard deviation are displayed.

Results regarding the Friedman-Rafsky test: This table shows detailed results about the test like the value of the statistic W.

Results regarding the minimum spanning tree: This table is displayed to give you a view of the minimum spanning tree. Four columns inform on the edges of the tree, nodes that the edge rely on, the weight (the distance between two nodes) and if the nodes come from the same sample.

Results regarding distance matrix: This table shows distances between each point of the two samples.

Example

A tutorial on Friedman-Rafsky is available on the XLSTAT Help Center:

https://www.xlstat.com/demo/rfy_en

References

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.

Testing for outliers

Grubbs test

Use this tool to test whether one or two outliers are present in a sample for which we assume that it is extracted from a population that follows a normal distribution.

In this section:

Description

Grubbs (1950, 1969, 1972) developed several tests in order to determine whether the greatest value or the lowest value (Grubbs test) are outliers, or, for the double Grubbs test, whether the two greatest values or the two lowest ones are outliers. This test assumes that the data corresponds to a sample extracted from a population that follows a normal distribution.

Detecting outliers

In statistics, an outlier is a value recorded for a given variable, that seems unusual and suspiciously lower or greater than the other observed values. One can distinguish two types of outliers:

- An outlier can simply be related to a reading error (on an measuring instrument), a keyboarding error, or a special event that disrupted the observed phenomenon to the point of making it incomparable to others. In such cases, you must either correct the outlier, if possible, or otherwise remove the observation to avoid that it disturbs the analyses that are planed (descriptive analysis, modeling, predicting).
- An outlier can also be due to an atypical event, but nevertheless known or interesting to study. For example, if we study the presence of certain bacteria in river water, you can have samples without bacteria, and other with aggregates with many bacteria. These data are of course important to keep. The models used should reflect that potential dispersion.

When there are outliers in the data, depending on the stage of the study, we must identify them, possibly with the aid of tests, flag them in the reports (in tables or on graphical representations), delete or use methods able to treat them as such.

To identify outliers, there are different approaches. For example, in classical linear regression, we can use the value of Cook's d values, or submit the standardized residuals to a Grubbs test to see if one or two values are abnormal. The classical Grubbs test can help identifying one outlier, while the double Grubbs test allows identifying two. It is not recommended to use these methods repeatedly on the same sample. However, it may be appropriate if you really suspect that there are more than two outliers.

Definitions

Let $x_1, x_2, \ldots, x_i, \ldots, x_n$, be a sample that is extracted from a population that we assume is following a normal distribution $N(\mu, s^2)$. Parameters μ and s^2 are respectively estimated by :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

We define:

$$x_{max} = \max_{i=1\ldots n}(x_i)$$

and

$$x_{min} = \min_{i=1\ldots n}(x_i)$$

The Grubbs test (for one outlier)

The statistics that used for the Grubbs test for one outlier are:

- For the one-sided left-tailed case: $G_{min} = \frac{\overline{x} - x_{min}}{s}$
- For the one-sided right-tailed case: $G_{\max} = \frac{x_{\max} - \overline{x}}{s}$
- For the two-sided case: $G = \max(G_{\min}, G_{\max})$

For the two-sided test, the null (H0) and alternative (Ha) hypotheses are given by:

- H0: The sample does not contain any outlier.
- Ha: The lowest or the greatest value is an outlier.

For the one-sided left-tailed test, the null (H0) and alternative (Ha) hypotheses are given by:

- H0: The sample does not contain any outlier.

- Ha: The lowest value is an outlier.

For the one-sided right-tailed test, the null (H0) and alternative (Ha) hypotheses are given by:

- H0: The sample does not contain any outlier.
- Ha: The greatest value is an outlier.

An approximation of the critical value G_{crit} giving the threshold above which, for a given significance level α (typically 5%), one must reject the null hypothesis is given by:

$$G_{crit}(n,\alpha) \approx \frac{(n-1)t_{n-2,1-\alpha/k}}{\sqrt{n-2+t^2_{n-2,1-\alpha/k}}}$$

where $t_{n-2,1-\alpha/k}$ is the value of the inverse of the Student cumulative distribution function at $1-\alpha/k$ with $n-2$ degrees of freedom, and where k equals n for one-sided tests and 2n for the two-sided test. We can compare this value to the G statistic computed for the sample, and deduce that one can keep H0 if G_{crit} is greater than G (or G_{min} or G_{max}) and reject it otherwise. From the G_{crit} approximation we can also deduce an approximation of the p-value that corresponds to G. XLSTAT displays all these results as well as the conclusion based on the significance level given by the user.

Double Grubbs test

For this test, we first sort up the xi observations. The statistics used for the double Grubbs test are given by:

- One-sided left-tailed test:

$$G2_{min} = \frac{Q_{min}}{(n-1)s}$$

with:

$$Q_{\min} = \sum_{i=3}^{n}(x_i - \overline{x}_3)^2, \overline{x}_3 = \frac{1}{n-2}\sum_{i=3}^{n} x_i$$

- One-sided right-tailed test:

$$G2_{max} = \frac{Q_{max}}{(n-1)s}$$

with:

$$Q_{max} = \sum_{i=1}^{n-2}(x_i - \overline{x}_{n-2})^2, \overline{x}_{n-2} = \frac{1}{n-2}\sum_{i=1}^{n-2} x_i$$

- Two-sided test:

$$G2_{minmax} = \max(G2_{\min}, G2_{\max})$$

For the two-sided test, the null (H0) and alternative (Ha) hypotheses are given by:

- H0: The sample does not contain any outlier.
- Ha: The two lowest or two greatest values are outliers.

For the one-sided left-tailed test, the null (H0) and alternative (Ha) hypotheses are given by:

- H0: The sample does not contain any outlier.
- Ha: The two lowest values are outliers.

For the one-sided right-tailed test, the null (H0) and alternative (Ha) hypotheses are given by:

- H0: The sample does not contain any outlier.
- Ha: The two greatest values are outliers.

Wilrich (2013) gives an approximation of the $G2_{crit}$ critical value above which one should reject H0 for a given significance level α. However XLSTAT gives an approximation based on Monte Carlo simulations. The default number of simulations is set to 1000000, which allows to obtain an accuracy that is higher than the ones available in original papers of Grubbs, and sufficient for any operational problem. Using the same set of simulations, XLSTAT gives the p-value that corresponds to the computed $G2$ statistic as well as the conclusion of the test taking into account the significance level given by the user.

Z-scores

Z-scores are displayed by XLSTAT to help you identify potential outliers. Z-scores correspond to the standardised sample:

$$z_i = \frac{x_i - \overline{x}}{s}, (i = 1, \ldots, n)$$

The problem with these scores is that once the acceptance interval is set (typically -1.96 and 1.96 for a 95% interval), any value that is outside is considered suspicious. However we know if we have 100 values, it is statistically normal to have 5 outside this interval. Furthermore, one can show that for a given n, the highest z-score is at most given by:

$$\max_{i=1\ldots n} z_i \leq \frac{n-1}{\sqrt{n}}$$

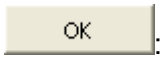
Iglewicz and Hoaglin (1993) recommend using a modified z-score in order to better identify outliers:

$$z_i = 0.6745\frac{x_i - \overline{x}}{MAD}, (i = 1, \ldots, n)$$

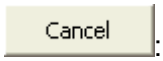
where the MAD is the *Median Absolute Deviation*. The acceptant interval is given by]-3.5 ; 3.5[whatever n.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data: Select the data on the Excel sheet. If you select several columns, XLSTAT considers column (or row in row mode) corresponds to a sample. If headers have been selected with the data, make sure the "Column labels" option is checked.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

You can choose the test to apply to your data:

- **Grubbs test**: Select this test to run a Grubbs test to identify one outlier.
- **Double Grubbs test**: Select this test to run a Grubbs test to identify two outliers.

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Significance level (%): Enter the significance level for the test (default value: 5%).

Iterations: Choose whether you want to apply the selected test data a limited number of times (default is 1), or if you want to let XLSTAT iterate until no more outlier is found.

Critical value / p-value: Enter the number of Monte Carlo simulations to perform to compute the critical value and the p-value, as well as the maximum time in seconds. This option is only available for the double Grubbs test.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Ignore missing data: Activate this option to ignore missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected samples.

Z-scores: Activate this option to calculate and display the z-scores and the corresponding graph. You can choose between the **modified z-scores** or standard **z-scores**. For z-scores you can choose which limits to display on the charts.

Results

Descriptive statistics: This table displays the descriptive statistics that correspond to the k samples.

The results correspond to the **Grubbs test** are then displayed. An interpretation of the test is provided if a single iteration of the test was requested, or if no observation was identified as being an outlier.

In case several iterations were required, also display a table showing, for each observation, the iteration in which it was removed from the sample.

The z-scores are then displayed if they have been requested.

Example

A tutorial showing how to use the Grubbs test is available on XLSTAT Help Center:

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

References

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.

Dixon test

Use this tool to test whether one or two outliers are present in a sample for which we assume that it is extracted from a population that follows a normal distribution.

In this section:

Description

The Dixon test (1950, 1951, 1953), which is actually divided into six tests depending on the chosen statistic and on the number of outliers to identify, was developed to help determine if the greatest value or lowest value of a sample, or the two largest values, or the two smallest ones can be considered as outliers. This test assumes that the data correspond to a sample extracted from a population that follows a normal distribution.

Detecting outliers

In statistics, an outlier is a value recorded for a given variable, that seems unusual and suspiciously lower or greater than the other observed values. One can distinguish two types of outliers:

- An outlier can simply be related to a reading error (on an measuring instrument), a keyboarding error, or a special event that disrupted the observed phenomenon to the point of making it incomparable to others. In such cases, you must either correct the outlier, if possible, or otherwise remove the observation to avoid that it disturbs the analyses that are planed (descriptive analysis, modeling, predicting).

- An outlier can also be due to an atypical event, but nevertheless known or interesting to study. For example, if we study the presence of certain bacteria in river water, you can have samples without bacteria, and other with aggregates with many bacteria. These data are of course important to keep. The models used should reflect that potential dispersion.

When there are outliers in the data, depending on the stage of the study, we must identify them, possibly with the aid of tests, flag them in the reports (in tables or on graphical representations), delete or use methods able to treat them as such.

To identify outliers, there are different approaches. For example, in classical linear regression, we can use the value of Cook's d values, or submit the standardized residuals to a Grubbs test

to see if one or two values are abnormal. The classical Grubbs test can help identifying one outlier, while the double Grubbs test allows identifying two. It is not recommended to use these methods repeatedly on the same sample. However, it may be appropriate if you really suspect that there are more than two outliers.

Definitions

Let $x_1, x_2, \ldots, x_i, \ldots, x_n$, be a sample of size n that is extracted from a population that we assume is following a normal distribution $N(\mu, \sigma^2)$. Parameters μ and σ^2 are respectively estimated by :

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

and

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n-1)}$$

We assume that the x_i's are sorted up.

Dixon test for one outlier

This test is used to determine if the largest or smallest value can be considered as being an outlier. This test assumes that the data corresponds to a sample coming from a population that follows a normal distribution.

The statistics used for the Dixon test and the corresponding ranges of number of observations they should be used with (Barnett and Lewis 1994 and Verma and Quiroz-Ruiz 2006) are:

- $R_{10} = \frac{x_n - x_{n-1}}{x_n - x_1}$, recommended for $3 \leq n \leq 100$, also named N7

- $R_{11} = \frac{x_n - x_{n-1}}{x_n - x_2}$, recommended for $4 \leq n \leq 100$, also named N9

- $R_{12} = \frac{x_n - x_{n-1}}{x_n - x_3}$, recommended for $5 \leq n \leq 100$, also named N10

These statistics are valid for testing whether the maximum value is an outlier. To identify if the minimum value is an outlier, simply sort the data in descending order and use the same statistics. If we want to identify if the minimum or the maximum value is an outlier, we calculate the statistics for the two alternatives (sort ascending or descending) and keep the largest value for the statistic.

For the two-sided test, the null (H0) and alternative (Ha) hypotheses are given by:

- H0: The sample does not contain any outlier.

- Ha: The lowest or the greatest value is an outlier.

For the one-sided left-tailed test, the null (H0) and alternative (Ha) hypotheses are given by:

- H0: The sample does not contain any outlier.
- Ha: The lowest value is an outlier.

For the one-sided right-tailed test, the null (H0) and alternative (Ha) hypotheses are given by:

- H0: The sample does not contain any outlier.
- Ha: The greatest value is an outlier.

Dixon test for two outliers

This test is used to determine if the two largest or two smallest values can be considered as being an outlier. This test assumes that the data corresponds to a sample coming from a population that follows a normal distribution.

The statistics used for the Dixon test and the corresponding ranges of number of observations they should be used with (Barnett and Lewis 1994 and Verma and Quiroz-Ruiz 2006) are:

- $R_{20} = \frac{x_n - x_{n-2}}{x_n - x_1}$, recommended for $4 \leq n \leq 100$, also named N11

- $R_{21} = \frac{x_n - x_{n-2}}{x_n - x_2}$, recommended for $5 \leq n \leq 100$, also named N12

- $R_{22} = \frac{x_n - x_{n-2}}{x_n - x_3}$, recommended for $6 \leq n \leq 100$, also named N13

These statistics are valid for testing whether the maximum value is an outlier. To identify if the minimum value is an outlier, simply sort the data in descending order and use the same statistics. If we want to identify if the minimum or the maximum value is an outlier, we calculate the statistics for the two alternatives (sort ascending or descending) and keep the largest value for the statistic.

For the two-sided test, the null (H0) and alternative (Ha) hypotheses are given by:

- H0: The sample does not contain any outlier.
- Ha: The two lowest or two greatest values are outliers.

For the one-sided left-tailed test, the null (H0) and alternative (Ha) hypotheses are given by:

- H0: The sample does not contain any outlier.
- Ha: The two lowest values are outliers.

For the one-sided right-tailed test, the null (H0) and alternative (Ha) hypotheses are given by:

- H0: The sample does not contain any outlier.
- Ha: The two greatest values are outliers.

Critical value and p-valuefor theDixon test

Literature provides more or less accurate approximations of the critical value beyond which, for a given significance level a, we cannot keep the null hypothesis. However XLSTAT provides an approximation of the critical values based on Monte Carlo simulations. The number of these approximations is by default set to 1000000, which provides more reliable than those provided in the historical articles. XLSTAT also provides on the basis of these simulations, a p-value and the conclusion of the test based on the significance level chosen by the user.

Z-scores

Z-scores are displayed by XLSTAT to help you identify potential outliers. Z-scores correspond to the standardised sample:

$$z_i = \frac{x_i - \overline{x}}{s} \quad (i = 1, \ldots, n)$$

The problem with these scores is that once the acceptance interval is set (typically -1.96 and 1.96 for a 95% interval), any value that is outside is considered suspicious. However we know if we have 100 values, it is statistically normal to have 5 of these outside this interval. Furthermore, one can show that for a given n, the highest z-score is at most given by:

$$\max_{i=1\ldots n} z_i \leq \frac{n-1}{\sqrt{n}}$$

Iglewicz and Hoaglin (1993) recommend using a modified z-score in order to better identify outliers:

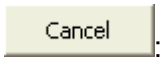
$$z_i = 0.6745 \frac{x_i - \overline{x}}{MAD} \quad (i = 1, \ldots n)$$

where the *MAD* is the *Median Absolute Deviation*. The acceptant interval is given by]-3.5 ; 3.5[whatever n.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data: Select the data on the Excel sheet. If you select several columns, XLSTAT considers column (or row in row mode) corresponds to a sample. If headers have been selected with the data, make sure the "Column labels" option is checked.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

You can choose the test to apply to your data:

- **User defined**: Choose this option to be able to select the statistic you want to use to identify outliers.
- **Automatic**: Choose this option to let XLSTAT choose the appropriate statistic, based on what is recommended in the literature (Böhrer, 2008).

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Significance level (%): Enter the significance level for the test (default value: 5%).

Iterations: Choose whether you want to apply the selected test data a limited number of times (default is 1), or if you want to let XLSTAT iterate until no more outlier is found.

Critical value / p-value: Enter the number of Monte Carlo simulations to perform to compute the critical value and the p-value as well as the maximum time in seconds.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Ignore missing data: Activate this option to ignore missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected samples.

Z-scores: Activate this option to calculate and display the z-scores and the corresponding graph. You can choose between the **modified z-scores** or standard **z-scores**. For z-scores you can choose which limits to display on the charts.

Results

Descriptive statistics: This table displays the descriptive statistics that correspond to the k samples.

The results correspond to the **Dixon test** are then displayed. An interpretation of the test is provided if a single iteration of the test was requested, or if no observation was identified as being an outlier.

In case several iterations were required, also display a table showing, for each observation, the iteration in which it was removed from the sample.

The z-scores are then displayed if they have been requested.

Example

A tutorial showing how to use the Dixon test is available on XLSTAT Help Center:

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

References

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.

Cochran’s C test

Use this tool to test whether there is an outlying variance among a series of k variances.

In this section:

Description

The Cochran’s C test (Cochran 1941) is one of the tests developed to identify and study the homogeneity of a series of variances (Bartlett's test, Brown- Forsythe, Levene or Hartley in particular). Cochran's test was developed to answer a specific question: Are the variances homogeneous or is the highest variance different from others? XLSTAT also offers two alternatives and uses the results of 't Lam (2010) for an extension of the balanced case to unbalanced cases, that also enables to perform two-sided tests.

Detecting outliers

In statistics, an outlier is a value recorded for a given variable, that seems unusual and suspiciously lower or greater than the other observed values. One can distinguish two types of outliers:

- An outlier can simply be related to a reading error (on a measuring instrument), a keyboarding error, or a special event that disrupted the observed phenomenon to the point of making it incomparable to others. In such cases, you must either correct the outlier, if possible, or otherwise remove the observation to avoid that it disturbs the analyses that are planed (descriptive analysis, modeling, predicting).

- An outlier can also be due to an atypical event, but nevertheless known or interesting to study. For example, if we study the presence of certain bacteria in river water, you can have samples without bacteria, and other with aggregates with many bacteria. These data are of course important to keep. The models used should reflect that potential dispersion.

When there are outliers in the data, depending on the stage of the study, we must identify them, possibly with the aid of tests, flag them in the reports (in tables or on graphical representations), delete or use methods able to treat them as such.

To identify outliers, there are different approaches. For example, in classical linear regression, we can use the value of Cook’s d values, or submit the standardized residuals to a Grubbs test

to see if one or two values are abnormal. The classical Grubbs test can help identifying one outlier, while the double Grubbs test allows identifying two. It is not recommended to use these methods repeatedly on the same sample. However, it may be appropriate if you really suspect that there are more than two outliers.

If the sample can be divided into sub-samples, we can look for changes from a sub-sample to another. Cochran's C test and Mandel's h and k statistics are part of the methods suitable for such studies.

Definitions

Let $x_{11}, x_{12}, \ldots, x_{1n1}, x_{12}, x_{22}, \ldots x_{2n2}, \ldots, x_{p1}, x_{p2}, \ldots, x_{pnp}$, be a sample that is divided into p groups (for example laboratories) of respective size $n_i (i = 1 \ldots p)$. Let $\overline{x}_i$ be the estimated mean for the i group, and let s_i^2 be the group i variance. We have:

$$\overline{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \overline{x_i})^2$$

It is assumed that the observations are identically distributed and follow a normal distribution.

Cochran's C test

The C_i statistic corresponding to group (or sub-sample) i $(i = 1 \ldots p)$ given by Cochran (1941) writes:

$$C_i = \frac{s_i^2}{\sum_{j=1}^{p} s_j^2}$$

and

$$C = \max_{i=1 \ldots p}(C_i)$$

is the statistic used for the test. The critical value corresponding to that statistic has abundantly been tabulated and various authors have given approximations (Wilrich, 2013). However, as noticed by 't Lam (2010), this statistic has several drawbacks:

- This test requires that the groups have identical sizes (balanced design),
- It is more difficult to study the minimum variance even if it is the true outlier (the left-tailed test is less efficient),
- Tables of critical values are limited and sometimes contain errors,
- The use of tables is not convenient.

For that reason, ‘t Lam suggests a generalization of the Cochran statistic for unbalanced designs, and a generalization test where the alternative hypothesis may be one or two-sided. The statistic for the group i is given by:

$$G_i = \frac{\nu_i S_i^2}{\sum_{j=1}^{p} \nu_j s_j^2} \text{ with } \nu_i = n_i - 1$$

For a significance level a, ‘t Lam gives the lower and upper critical values for this statistic:

$$G_{LL}(i) = \left[1 + \frac{(\nu_{total}/\nu_i) - 1}{F^{-1}(\delta/p, \nu_i, \nu_{total} - \nu_i)}\right]^{-1} \quad (1)$$

and

$$G_{UL}(i) = \left[1 + \frac{(\nu_{total}/\nu_i) - 1}{F^{-1}(1 - \delta/p, \nu_i, \nu_{total} - \nu_i)}\right]^{-1} \quad (2)$$

with $\nu_{total} = (\sum_{i=1}^{p} \nu_i) - p$, and $\delta = \alpha$ for a one-sided test and $\delta = \alpha/2$ for a two-sided test, and F^{-1} is the inverse Fisher cumulative distribution function.

For the two-sided test, the null (H_0) and alternative (H_a) hypotheses are given by:

- H_0: The variances are homogeneous.
- H_a : One of the variances is lower than the others.

For the one-sided left-tailed test, the null (H_0) and alternative (H_a) hypotheses are given by:

- H_0 : The variances are homogeneous.
- H_a : At least one of the variances is lower than the others.

For the one-sided right-tailed test, the null (H_0) and alternative (H_a) hypotheses are given by:

- H_0 : The variances are homogeneous.
- H_a : At least one of the variances is greater than the others.

Under a two-sided test, to identify the potentially outlying variance we compute:

$$Gmin = \min_{i=1...k}(G_i) \text{ and } Gmax = \max_{i=1...k}(G_i)$$

Then, if one or two statistics are not within the critical range given by (1) and (2), the p-values associated with the two statistics are calculated. We identify the abnormal variance as the one that corresponds to the lowest p-value.

Z-scores

Z-scores are displayed by XLSTAT to help you identify potential outliers. Z-scores correspond to the standardised sample:

$$z_i = \frac{x_i - \overline{x}}{s} \; (i = 1, ...n)$$

The problem with these scores is that once the acceptance interval is set (typically -1.96 and 1.96 for a 95% interval), any value that is outside is considered suspicious. However we know if we have 100 values, it is statistically normal to have 5 outside this interval. Furthermore, one can show that for a given n, the highest z-score is at most given by:

$$\max_{i=1...n} z_i \leq \frac{n-1}{\sqrt{n}}$$

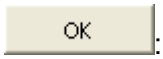
Iglewicz and Hoaglin (1993) recommend using a modified z-score in order to better identify outliers:

$$z_i = 0.6745 \frac{x_i - \overline{x}}{MAD} \; (i = 1, ...n)$$

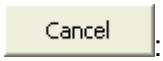
where the MAD is the *Median Absolute Deviation*. The acceptance interval is given by $]-3.5; 3.5[$ whatever n.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data: If the format of the selected data is "one column per variable", select the data for the various samples in the Excel worksheet. If the format of the selected data is "one column per group", select the columns of data corresponding to the various groups. If the format of the selected data is "variances", select the variances of each group.

Group identifiers / Group size: If the format of the selected data is "one column per variable", select the data identifying the groups to which the selected data values correspond. If the format of the selected data is "Variances" you need to enter the group size (balanced design) or select the group sizes (unbalanced design).

Data format: Select the data format.

- **One column/row per group**: Activate this option to select one column (or row in row mode) per group.
- **One column/row per variable**: Activate this option for XLSTAT to carry out as many tests as there are columns/rows, given that each column/row must contain the same number of rows/columns and that a sample identifier which enables each observation to be assigned to a sample must also be selected.
- **Variances**: Activate this option if your data correspond to variances. In that case you need to define the group size (balanced design) or select the group sizes (unbalanced design).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

You can choose the test to apply to your data:

- **Cochran's C (balanced)**: Select this option if the design is balanced and you want to perform a one-sided test.
- **'t Lam 's G (unbalanced)**: Select this option if the design is unbalanced and/or you want to perform a two-sided test.

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see description).

Significance level (%): Enter the significance level for the test (default value: 5%).

Iterations: Choose whether you want to apply the selected test data a limited number of times (default is 1), or if you want to let XLSTAT iterate until no more outlier is found.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected samples.

Z-scores: Activate this option to calculate and display the z-scores and the corresponding graph. You can choose between the **modified z-scores** or standard **z-scores**. For z-scores you can choose which limits to display on the charts.

Results

Descriptive statistics: This table displays the descriptive statistics that correspond to the groups.

The results correspond to the **Cochran's C test** are then displayed. An interpretation of the test is provided if a single iteration of the test was requested, or if no observation was identified as being an outlier.

In case several iterations were required, also display a table showing, for each observation, the iteration in which it was removed from the sample.

The z-scores are then displayed if they have been requested.

Example

A tutorial showing how to use the Cochran test is available on XLSTAT Help Center:

http://www.xlstat.com/demo-cochran-c.htm

References

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.

Mandel’s h and k statistics

Use this tool to calculate the h and k Mandel's statistics to identify potential outliers in a sample.

In this section:

Description

The Mandel’s h and k statistics (1985, 1991) have been developed to help identifying outliers during inter-laboratories studies.

Detecting outliers

In statistics, an outlier is a value recorded for a given variable, that seems unusual and suspiciously lower or greater than the other observed values. One can distinguish two types of outliers:

- An outlier can simply be related to a reading error (on an measuring instrument), a keyboarding error, or a special event that disrupted the observed phenomenon to the point of making it incomparable to others. In such cases, you must either correct the outlier, if possible, or otherwise remove the observation to avoid that it disturbs the analyses that are planed (descriptive analysis, modeling, predicting).

- An outlier can also be due to an atypical event, but nevertheless known or interesting to study. For example, if we study the presence of certain bacteria in river water, you can have samples without bacteria, and other with aggregates with many bacteria. These data are of course important to keep. The models used should reflect that potential dispersion.

When there are outliers in the data, depending on the stage of the study, we must identify them, possibly with the aid of tests, flag them in the reports (in tables or on graphical representations), delete or use methods able to treat them as such.

To identify outliers, there are different approaches. For example, in classical linear regression, we can use the value of Cook’s D values, or submit the standardized residuals to a Grubbs test to see if one or two values are abnormal. The classical Grubbs test can help identifying one outlier, while the double Grubbs test allows identifying two. It is not recommended to use these

methods repeatedly on the same sample. However, it may be appropriate if you really suspect that there are more than two outliers.

If the sample can be divided into sub-samples, we can look for changes from a sub-sample to another. The test Cochran's C test and the Mandel's h and k statistics are part of the methods suitable for such studies.

Definitions

Let $x_{11}, x_{12}, \ldots, x_{1n_1}, x_{12}, x_{22}, \ldots, x_{2n_2}, \ldots, x_{p1}, x_{p2}, \ldots, x_{pn_p}$, be a sample of that we distinguish for their belonging to p groups (for example laboratories) of respective size n_i $(i = 1, \ldots p)$. Let $\overline{x}_i$ be the estimated mean for the i group, and let s_i^2 be the group i variance. We have:

$$\overline{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$$

and

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n} \left(x_{ij} - \overline{x}_i\right)^2$$

It is assumed that the observations are identically distributed and follow a normal distribution.

Mandel's h statistic

Mandel's h_i statistic for group i, $(i = 1, \ldots, p)$ is given by:

$$h_i = \frac{\overline{x}_i - \overline{\overline{x}}}{s}$$

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 provides h_i statistics for each group. To identify groups for which the mean is potentially abnormal, we can calculate the critical values and confidence intervals for a given level of significance α around statistic h (Wilrich, 2013). The critical value is given by:

$$h_{crit}(p, \alpha) = \frac{(p-1)t_{p-2,1-\alpha/2}}{\sqrt{p(p-2+t_{p-2,1-\alpha/2}^2)}}$$

where t corresponds to the quantile of the Student distribution for $1 - \alpha/2$ and $p - 2$ degrees of freedom.

The confidence interval (two-sided) of size $100(1-\alpha)\%$ around h_i is given by $\left]-h_{i,crit};h_{i,crit}\right[$. XLSTAT displays the critical value on the chart of the h_i if the n_i are constant.

Mandel's k statistic

Mandel's k_i for group i $(i=1,\ldots,p)$ is given by:

$$k_i = \frac{s_i}{\tilde{s}}$$

with

$$s_i = \sqrt{\frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(x_{ij}-\overline{x}_i\right)^2} \text{ et } \tilde{s} = \sqrt{\frac{1}{p}\sum_{i=1}^{p}s_i^2}$$

XLSTAT provides k_i statistics for each group. To identify groups for which the variance is potentially abnormal, we can calculate the critical values and confidence intervals for a given level of significance α around statistic k (Wilrich, 2013). The critical value is given by:

$$k_{crit}(n,\alpha) = \sqrt{p(1+(p-1)F^{-1}_{1-\alpha,(p-1)(n-1),(n-1)})}$$

where $F^{-1}_{1-\alpha,v_1,v_2}$ is the value of the inverse cumulative distribution function of the Fisher distribution for probability $1-\alpha$ with v_1 and v_2 degrees of freedom.

The confidence interval (one-sided) of size $100(1-\alpha)\%$ around k_i is given by $\left[0;k_{i,crit}\right]$. XLSTAT displays the critical value on the chart of the k_i if the n_i are constant.

Z-scores

Z-scores are displayed by XLSTAT to help you identify potential outliers. Z-scores correspond to the standardised sample:

$$z_i = \frac{x_i-\overline{x}}{s}(i=1,\ldots,n)$$

The problem with these scores is that once the acceptance interval is set (typically -1.96 and 1.96 for a 95% interval), any value that is outside is considered suspicious. However we know if we have 100 values, it is statistically normal to have 5 outside this interval. Furthermore, one can show that for a given n, the highest z-score is at most given by:

$$\max_{i=1\ldots n} z_i \leq \frac{n-1}{\sqrt{n}}$$

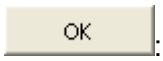
Iglewicz and Hoaglin (1993) recommend using a modified z-score in order to better identify outliers:

$$z_i = 0.6745 \frac{x_i - \overline{x}}{MAD} \quad (i = 1, \ldots n)$$

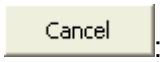
where the MAD is the *Median Absolute Deviation*. The acceptant interval is given by $]-3.5; 3.5[$ whatever n.

Dialog box

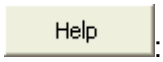
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data: If the format of the selected data is "one column per variable", select the data for the various samples in the Excel worksheet. If the format of the selected data is "one column per group", select the columns of data corresponding to the various groups. If the format of the selected data is "variances", select the variances of each group. If the format of the selected data is "means", select the means of each group.

Group identifiers / Group size: If the format of the selected data is "one column per variable", select the data identifying the groups to which the selected data values correspond. If the format of the selected data is "Variances" or "Means" you need to enter the group size (balanced design).

Data format: Select the data format.

- **One column/row per group**: Activate this option to select one column (or row in row mode) per group.

- **One column/row per variable**: Activate this option for XLSTAT to carry out as many tests as there are columns/rows, given that each column/row must contain the same number of rows/columns and that a sample identifier which enables each observation to be assigned to a sample must also be selected.
- **Variances**: Activate this option if your data correspond to variances. In that case you need to define the sample size (balanced design).
- **Means**: Activate this option if your data correspond to means, In that case you need to define the sample size (balanced design).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row (column mode) or first column (rows mode) of the selected data contain labels.

You can choose the test to apply to your data:

- **Mandel's h statistic**: choisissez cet option pour calculer la statistique h de Mandel.
- **Mandel's k statistic**: choisissez cet option pour calculer la statistique k de Mandel.

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test.

Significance level (%): Enter the significance level for the test (default value: 5%).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected samples.

Z-scores: Activate this option to calculate and display the z-scores and the corresponding graph. You can choose between the **modified z-scores** or standard **z-scores**. For z-scores you can choose which limits to display on the charts.

Results

Descriptive statistics: This table displays the descriptive statistics that correspond to the groups.

The results correspond to the **Mandel's statistics** are then displayed.

The z-scores are then displayed if they have been requested.

Example

A tutorial showing how to compute the Mandel's statistics is available on XLSTAT Help Center:

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

References

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

Use Easy Fit to test and compare different predictive models for the same dataset. Depending on the type of the dependent and explanatory variables (quantitative or qualitative), various models are proposed. The Easy Predict function can be then used to make predictions using the previously fitted models.

In this section:

Description

Introduction

When trying to predict the values of a quantitative Y variable, we talk about **regression**, whereas we talk about **classification** when the Y variable to predict is qualitative. XLSTAT offers several regression and classification learning models. The Easy Fit function was developed to address the following two main issues:

- When using a predictive model within XLSTAT, a lot of results are available (by default or optional). This is essential for some experts but not for users who simply want to know if the fitted model is working well and therefore they do not need all the provided results.
- Historically in XLSTAT, if you want to compare several models for the same dataset, you need to separately run the models thus configure different XLSTAT dialog boxes. This requires several steps which may take some time.

The Easy Fit feature provides solutions to the two previous points. It allows, depending on the nature of the problem (regression or classification), to quickly generate several different models using the same dataset. The results of the fitted models are very synthetic so that the user can determine at one glance which is the best model.

The quality of the models is assessed using indicators. These indicators are calculated on a validation sample containing 20% of randomly selected observations.

Available regression models

Easy Fit offers the following regression models (you can click on the different methods to access the associated help document):

- If the X explanatory variables are only **quantitative**:
    - Linear regression
    - Regression random forests
    - K Nearest Neighbors
    - Support Vector Machine
    - LASSO Regression
    - Extreme Gradient Boosting
- If the X explanatory variables are only **qualitative**:
    - Analysis of variance: ANOVA
    - Regression random forests
    - K Nearest Neighbors
    - Support Vector Machine
    - LASSO Regression
    - Extreme Gradient Boosting
- If some X explanatory variables are **quantitative and other qualitative**:
    - Analysis of covariance analysis: ANCOVA
    - Regression random forests
    - K Nearest Neighbors
    - Support Vector Machine
    - LASSO Regression
    - Extreme Gradient Boosting

Available classification models

Easy Fit offers the following classification models (you can click on the different methods to access the associated help document): - Logistic regression - Classification random forests - K Nearest Neighbors - Support Vector Machine - LASSO Regression - Extreme Gradient Boosting

Presentation of results

At the beginning of each Easy Fit results sheet, you will find a summary table. This table contains the quality indicators of the models. This table allows you to quickly see which is the best model performed.

If Easy Fit is used for a regression, the summary table contains the following indicators: MAE, MSE, R^2, adjusted R^2, AIC and SBC. Therefore for a classification, the indicators are: Accuracy, Precision, Recall, Correct classification number, Misclassification number and F-score.

You will also find the synthetic results of each model. For more details, please refer to the related help documents. For each template you will find at the beginning of the results two buttons.

 : This button allows you to automatically open the pre-filled dialog box of the complete method. This will give you access to all possible options and more results given by the chosen method.

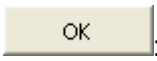
 : This button opens the Easy Predict function dialog box to make predictions for new observations.

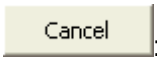
Easy Predict

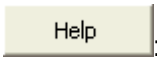
The purpose of the Easy Predict function is to predict the values of new observations which were not used for model learning or validation. Use the Easy Predict button (see above) to open a new dialog box and select the new observations to predict using the model generated with Easy Fit. The output of Easy Predict contains the predictions associated with these new observations.

Dialog box

The dialog box is divided into several options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

 : Click on this button to start the computations.

 : Click on this button to close the dialog box without doing any computation.

 : Click this on button to display the help.

 : Click on this button to reload the default options.

 : Click this on button to delete the data selections.

  : Click on these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

Type of the Y variable to predict: Choose the type of Y variable you want to predict. If Y is quantitative then you will be able to choose between regression methods and if Y is qualitative then you will be able to choose between classification methods. Then select the variable to be predicted in the associated field. If the data header has been selected, check that the "Variable labels" option has been activated.

Type of explanatory X variables: Choose the type of explanatory (or predictive) X variables to use in your model. You can use only quantitative variables, only qualitative variables, or both. Then select your explanatory variables in the associated fields. If the variable header has been selected, check that the "Variable labels" option has been activated.

Display results in:

- **New worksheet**: Activate this option to display the results in a new worksheet of the active workbook. In this case, you can give the result sheet a name. If you do not specify a name, a default name will be created.
- **New workbook**: Activate this option to display the results in a new workbook. In this case, you can give the result sheet a name. If you do not specify a name, a default name will be created.
- **Existing cell**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Variable labels: Activate this option if the first row of the data selections includes a header.

Select the methods you want to use: Depending on the types of variables used, different methods are proposed (see section Description for a list).

Descriptive statistics: Activate this option to display descriptive statistics on the different variables used.

Results

Descriptive statistics: The tables of descriptive statistics show basic statistics for all the selected variables such as the number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables. For qualitative variables the names of the various categories are displayed together with their respective frequencies.

Summary table: This table presents the adjustment coefficients for assessing the quality of the model. Adjustment coefficients are calculated on the validation sample. The number of misclassified observations is mostly used for classification models and the mean squared errors for regression models. For both, the goal is to get the lower possible values.

Summary results by method: For each method the most important results of the method are displayed. For a better understanding of the various results, you can access the help section of the related methods. All links to the help sections are available in the [Description] section (#description).

Example

An example of using the Easy Fit function and the Easy Predict function is available on the XLSTAT Help Center at:

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

Mathematical tools

Probability calculator

Use this tool to compute for a given distribution function, the density function, the cumulative distribution function, or the inverse cumulative distribution function.

In this section:

Description

A probability distribution $\mathbb{P}$ is a special type of function, that is named a *measure* in mathematics. It allows to relate (thanks to what mathematicians name a *map*) events (for example, the occurrence of a 2 when throwing a die), to their probability. Mathematicians distinguish the Ω set which represents all the possible outcomes when running a random experiment, from the set $\mathcal{A}$ which corresponds to the subset of events that are measurable. In the case of a discrete distribution (where only integers are manipulated), Ω and $\mathcal{A}$ are identical. The triplet $(\Omega, \mathcal{A}, \mathbb{P})$ is named the *measurable space*.

A large number of probability distributions have been developed to describe particular situations where randomness occurs. Randomness is a representation of ignorance or imperfect knowledge. When you roll a die, a perfect knowledge of the starting movement, the environment, the forces involved, ..., would allow you knowing when, where, and in what position the die would stop. Nevertheless, it is so complex that it is preferable to estimate that each face has a certain probability of being the result of the throw and that the occurrence of an event is random.

A random variable is also a function that maps an event to a real. For example, we can match the tossing of a coin with 0 if the coin lands on head, and 1 if the coin lands on tail. While this is perfectly arbitrary in the case of a coin, it may be more natural if one counts the number of people in a queue at the post office, or if one measures the air temperature. In this case, we will match the event "there are 10 people queuing" with the number 10, or "the temperature is 59.8°F" with the real 59.8.

In the case of discrete variables, each event has a non-zero probability as long as it is possible. In the case of continuous variables, each (possible) event has a non-zero probability of

occurring, but it is so small that only the probability distribution makes it possible to measure it. While the measure does not mean much by itself, it allows to compare the relative chances to occur of two events. It also allows, through a sum (an integral) to give the probability of a series of events (described by an interval) to occur. Thus, for example, if a measurement of the temperature is made and the measurement is known to be error-prone and if we assume the measurement follows a normal distribution (the famous distribution with a bell shape), the event "it is 60°F", has virtually a null probability to occur, while on the other hand we will be able to give the probability that we measure a temperature between 59.5°F and 60.5°F. The mathematical tool for calculating this probability is called the **cumulative distribution function** (CDF).

In probability theory, the probability distribution is described by the cumulative distribution function which can itself depend on different parameters. The simplest law is Bernouilli's law where two events are possible (typically flipping of a coin). Its unique parameter is the probability p_0 for the coin to fall on head and if the coin is not loaded, we have $p_0 = 0.5$. For the normal distribution, an essential distribution in statistics, the parameters are the mean μ and the variance σ^2.

For a continuous random variable, the cumulative distribution function is an increasing function that takes values in the interval $[0; 1]$.

Two other functions are commonly used:

- The probability density function: This function is the function to integrate to calculate the cumulative distribution function, which is valid for the case of variables with density (which is true for all distributions proposed by XLSTAT).
- The inverse cumulative distribution function: This function allows, for a given probability p, to obtain the value x of the random variable such that the distribution function in x is p.

The probability calculator allows to calculate, for all distributions proposed by XLSTAT, the probability density function, the cumulative distribution function and the inverse cumulative distribution function.

Dialog box

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

Distribution: Select the distribution for which you want to make computations. You can then enter the value of the parameters of the distribution. A description of the different distributions can be found in this section of the help.

Calculate: Select the function that you want to compute (probability distribution function, cumulative distribution function, or inverse cumulative distribution function), then enter the point at which you want to compute it. For the cumulative distribution function F, you can choose among:

- $< a$: The result corresponds to $F(a)$
- $> a$: The result corresponds to $1 - F(a)$
- $a << b$: The result corresponds to $(F(b) - F(a)$
- $< a \quad b <$: The result corresponds to $1 - (F(b) - F(a))$

Calculate: Click this button to display the result in the *Results* part of the dialog box.

Display results in:

- **Range**: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.
- **Sheet**: Activate this option to display the results in a new worksheet of the active workbook.
- **Workbook**: Activate this option to display the results in a new workbook.

Display the report header: Deactivate this option if you do not want to display the report header.

Results: The results are displayed in this part of the dialog box. You may copy them (Ctrl C) to paste it another software.

Clear: Click this button to clear the results that are displayed in the Results part of the dialog box.

Results

The results displayed correspond to the probability distribution function and to the value computed for the selected function (CDF, PDF or inverse CDF) at a specific point.

Example

An example showing how to use the probability calculator is available at:

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

References

Krishnamoorthy K. (2015). Handbook of Statistical Distributions with Applications. Chapman and Hall/CRC.

Matrix operations

Use this tool to perform operations (addition, subtraction or product) on matrices.

In this section:

Description

This tool allows to compute an inverse matrix, a transpose matrix and perform operations between two matrices A and B. The user can choose to use the original matrix or transform it prior to the operation. Two options are available:

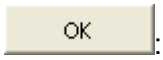
- Use the transpose matrix, noted as A' for the A matrix.
- Use the inverse matrix, noted as A^(-1) for the A matrix.

If a second matrix B is selected then different matrix operations are available:

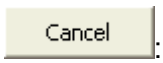
- Addition: matrices A and B must be of the same size (same number of rows and same number of columns).
- Subtraction: matrices A and B must be of the same size (same number of rows and same number of columns).
- Product: the number of rows in matrix B must be equal to the number of columns in matrix A.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select

General tab:

Matrix A: Select the data corresponding to matrix A. If a column header has been selected, check that the "Variable labels" option has been activated.

- **Transpose: A'**: Select this option if you want to simply compute the transpose of matrix A or use the transpose of A for the operation between matrices A and B.
- **Inverse: A^(-1)**: Select this option if you want to simply compute the inverse of matrix A or use the inverse of A for the operation between matrices A and B. If the matrix is non-invertible, the calculations will stop.

Matrix B: Select this option if you want to perform an operation between two matrices. In this case, select the data corresponding to matrix B. If a column header has been selected, check that the "Variable labels" option has been activated.

- **Transposed: B'**: Select this option if you want to use the transpose of B instead of B.
- **Inverse: B^(-1)**: Select this option if you want to use the inverse of B instead of B. If the matrix is non-invertible, the calculations will stop.

Operation: Select the desired operation to perform between matrices A and B: addition, substraction or product (see the description section for further details).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the data selections include a header.

Display the report header: Deactivate this option if you do not want to display the report header.

Result: Based on the selected options in the dialog box, the final operation to be performed is displayed here.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Ignore missing data: If you choose this option, missing data will be retained. All operations involving missing data will return missing data.

Results

Result: The main output of this tool displays the computed matrix based on the chosen operation.

Example

A tutorial which explains how to use the Matrix Operations tool is available here:

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

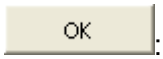
Tools

DataFlagger

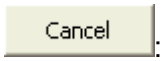
Use DataFlagger to show up the values within or outside a given interval, or which are equal to certain values.

In this section:

Dialog box



: Click this button to start flagging the data.



: Click this button to close the dialog box without doing any change.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

Data: Select the data in the Excel worksheet.

Flag a value or a text: Activate this option is you want to identify or show up a value or a series of values in the selected range.

- **Value or text**: Choose this option to find and flag a single value or a character string.
- **List values or texts**: Choose this option to find and flag a series of values or texts. You must then select the series of values or texts in question in an Excel worksheet.

Flag an interval: Activate this option is you want to identify or show up values within or outside an interval. You then have to define the interval.

- **Inside**: Choose this option to find and flag values within an interval. Afterwards choose the boundary types (open or closed) for the interval, then enter the values of the

boundaries.

- **Outside**: Choose this option to find and flag values outside an interval. Afterwards choose the boundary types (open or closed) for the interval, then enter the values of the boundaries.

Font: Use the following options to change the font of the values obeying the flagging rules.

- **Style**: Choose the font style
- **Size**: Choose the font size
- **Color**: Choose the font color

Cell: Use the following option to change the background color of the cell.

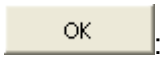
- **Color**: Choose the cell color

Min/Max Search

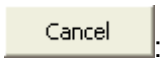
Use this tool to locate the minimum and/or maximum values in a range of values. If the minimum value is encountered several times, XLSTAT makes a multiple selection of the minimum values enabling you afterwards to browse between them simply using the "Enter" key.

In this section:

Dialog box



: Click this button to start the search.



: Click this button to close the dialog box without doing any search.

: Click this button to display the help.

Data: Select the data in the Excel worksheet.

Find the minimum: Activate this option to make XLSTAT look for the minimum value(s) in the selection. If the "Multiple selection" option is activated and several minimum values are found, they will all be selected and you can navigate between them using the "Enter" key.

Find the maximum: Activate this option to make XLSTAT look for the maximum value(s) in the selection. If the "Multiple selection" option is activated and several maximum values are found, they will all be selected and you can navigate between them using the "Enter" key.

Multiple selection: Activate this option to enable multiple occurrences of the minimum and/or maximum values to be selected at the same time.

Remove text values in a selection

Use this tool to remove text values in a data set that is expected to contain only numerical data. This tool is useful if you are importing data from a format that generates empty cells with a text format in Excel.

In this section:

Dialog box

OK: Click this button to start removing the text values.

Cancel: Click this button to close the dialog box without doing any change.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

Data: Select the data in the Excel worksheet.

Clean only the cells with empty strings: Activate this option if you want to only clean the cells that correspond to empty strings.

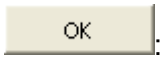
Upper and lower case

Use this tool to trim spaces to change the upper or lower case settings of text data.

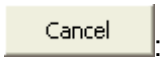
In this section:

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: click this button to change the way you load the data in XLSTAT. If the button has a mouse icon, XLSTAT allows you to make a mouse selection of your data. If the button has a list icon, XLSTAT allows you to select the columns named by their first item. If the button has an orange paper sheet icon, additional buttons with a question mark appeared  . These allow to select a file and its reading parameters (see Import data file).

General tab:

Data: Select data on the worksheet, in a list or from a file.

Lower case: Activate this option to convert text data to lower case.

- **First letter upper case**: Activate this option to capitalize the first letter of each word.
- **First word only**: Activate this option to only capitalize the first letter of the first word.

Upper case: Activate this option to convert text data to upper case.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Column labels: Check this option if the first row of the selected data contains a label.

Display the report header: Deactivate this option if you do not want to display the report header.

Results

The results are displayed at the desired location. The table contains the strings processed by the methods chosen.

Sheets management

Use this tool to manage the sheets contained in the open Excel workbooks.

In this section:

Dialog box

When you start this tool, it displays a dialog box that lists all the sheets contained in all the workbooks, whether they are hidden or not.

Activate: Click this button to go to the first sheet that is selected.

Unhide: Click this button to unhide all the selected sheets.

Hide: Click this button to hide all the selected sheets.

Delete: Click this button to delete all the selected sheets. Warning: deleting hidden sheets is irreversible.

Cancel: Click this button to close the dialog box.

Help: Click this button to display help.

Delete hidden sheets

Use this tool to delete the hidden sheets generated by XLSTAT or other applications. XLSTAT generates hidden sheets to create certain charts. This tool is used to choose which hidden sheets are to be deleted and which kept.

In this section:

Dialog box

Hidden sheets: The list of hidden sheets is displayed. Select the hidden sheets you want to delete.

All: Click this button to select all the sheets in the list.

None: Click this button to deselect all the sheets in the list.

Delete: Click this button to delete all the selected sheets. Warning: deleting hidden sheets is irreversible.

Cancel: Click this button to close the dialog box.

Help: Click this button to display help.

Unhide hidden sheets

Use this tool to unhide the hidden sheets generated by XLSTAT or other applications. XLSTAT generates hidden sheets to create certain charts.

In this section:

Dialog box

Hidden sheets: The list of hidden sheets is displayed. Select the hidden sheets you want to unhide.

All: Click this button to select all the sheets in the list.

None: Click this button to deselect all the sheets in the list.

Unhide: Click this button to unhide all the selected sheets.

Cancel: Click this button to close the dialog box.

Help: Click this button to display help.

Export to GIF/JPG/PNG/TIF

Use this tool to export a table, a chart, or any selected object on an Excel sheet to a GIF, JPG, PNG ou TIF file.

In this section:

Dialog box

Save: Click this button to save the selected object to a file.

Cancel: Click this button to close the dialog box.

Help: Click this button to display the help.

 : Click this button to reload the default options.

Format: Choose the graphic format of the file.

File name: Enter the name of the file to which the image should be saved, or select the file in a folder.

Resize: Activate this option to modify the size of the graphic before saving it to a file.

- **Width**: Enter the value in points of the graphic's width;
- **Height**: Enter the value in points of the graphic's height.

Display the grid: Activate this option if you want that while generating the file, XLSTAT keeps the gridlines that separate the cells. This option is only active when cells or tables are selected.

Add comments

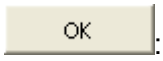
Use this tool to create or add comments to spreadsheet cells, using content that is available in other cells.

In this section:

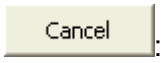
Description

Comments in Excel cells are interesting because they make it optional to view certain information, for example explanatory. However, loading them is not necessarily easy. Thanks to this XLSTAT tool you can easily create or modify comments..

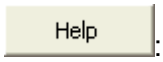
Dialog box



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.

Comments: Select the comment(s) you want to add to cells. The layout of the comments should match the layout of the cells you want to add these comments to.

Cells to comment: Select the cell or cells to which you want to add comments. The layout of the comments should match the layout of the cells you want to add these comments to.

Merge: Activate this option so that if a cell already has a comment, the new text is added to the existing comment.

Example

An example on how to use the tool is available on XLSTAT Help Center:

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

Sensory data analysis

External Preference Mapping (PREFMAP)

Use this method to model and represent graphically the preference of assessors for a series of objects depending on objective criteria or linear combinations of criteria.

In this section:

Description

External preference mapping (PREFMAP) is used to display on the same chart (in two of three dimensions) objects and indications showing the preference levels of assessors (in general, consumers) in certain points in the representation space. The preference level is represented on the preference map in the form of vectors, ideal or anti-ideal points, or isopreference curves depending on the type of model chosen.

These models are themselves constructed from objective data (for example physico-chemical descriptors, or scores provided by experts on well-determined criteria) which enable the position of the assessors and the products to be interpreted according to objective criteria.

If there are only two or three objective criteria, the axes of the representation space are defined by the criteria themselves (possibly standardized to avoid the effects of scale). On the other hand, if the number of descriptors is higher, a method for reducing the number of dimensions must be used. In general, PCA is used. Nevertheless, it is also possible to use factorial analysis if it is suspected that underlying factors are present, or MDS (multidimensional scaling) if the initial data are the distances between the products. If the descriptors used by the experts are qualitative variables, a PCA can be used to create a 2- or 3-dimensional space.

The PREFMAP can be used to answer the following questions:

- How is the product positioned with respect to competitive products?

- What is the nearest competitive product to a given product?

- What type of consumer prefers a product?

- Why are certain products preferred?

- How can I reposition a product so that it is again more preferred by its core target?

- What new products might it be relevant to create?

Preference models

To model the preferences of assessors depending on objective criteria or a combination of objective criteria (if a PCA has enabled a 2- or 3-dimensional space to be created) four models have been proposed within the framework of PREFMAP. For a given assessor, if we designate y_i to be their preference for product i, and X1, X2, …, Xp to be the p criteria or combinations of criteria (in general p=2) describing product i, the models are:

- **Vector**: $y_i = a_0 + \sum_{j=1}^{p} a_j x_{ij}$
- **Circular**: $y_i = a_0 + \sum_{j=1}^{p} a_j x_{ij} + b \sum_{j=1}^{p} x_{ij}^2$
- **Elliptic**: $y_i = a_0 + \sum_{j=1}^{p} a_j x_{ij} + \sum_{j=1}^{p} b_j x_{ij}^2$
- **Quadratic**: $y_i = a_0 + \sum_{j=1}^{p} a_j x_{ij} + \sum_{j=1}^{p} b_j x_{ij}^2 + \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} c_{jk} x_{ij} x_{ik}$

The coefficients a,b,c are estimated by multiple linear regression. It will be noted that the models are classified from the simplest to the most complex. XLSTAT lets you either chose one model to use for all assessors, or choose a model giving the best result as regards the p-value of Fisher's F for a particular assessor or the p-value of the F-ratio test. In other words, you can choose a model which is both parsimonious and powerful at the same time.

The **vector model** represents individuals on the sensorial map in the form of vectors. The size of the vectors is a function of the R² of the model: the longer the vector, the better the corresponding model. The preference of the assessor will be stronger the further you are in the direction indicated by the vector. The interpretation of the preference can be done by projecting the different products on the vectors (product preference). The disadvantage of the vector model is that it neglects the fact that for certain criteria, like the saltiness or temperature for example, there can be an increase of preference to an optimum value then a decrease.

The **circular model** takes into account this concept of optimum. If the surface area for the model has a maximum in terms of preference (this happens if the b coefficient is estimated negative), this is known as the ideal point. If the surface area for the model has a minimum in terms of preference (this happens if the b coefficient is estimated positive), this is known as the anti-ideal point. With the circular model, circular lines of isopreference can be drawn around the ideal or anti-ideal points.

The **elliptical model** is more flexible, as it takes the effect of scale into account better. The disadvantage of this model is that there is not always an optimum: as with the circular model, it can generate an ideal point or an anti-ideal point if all the b_j coefficients have the same sign, but we may also obtain a saddle point (in the form of a surface shaped like a horse's saddle) if all the b_j coefficients do not have the same sign. The saddle point cannot easily be interpreted. It corresponds only to an area where the preference is less sensitive to variations.

Lastly, the **quadratic model** takes more complex preference structures into account, as it includes interaction terms. As with the elliptical model we can obtain an ideal, an anti-ideal, or a saddle point.

Preference map

The preference map is a summary view of three types of elements:

The assessors (or groups of assessors if a classification of assessors has been carried out beforehand) represented in the corresponding model by a vector, an ideal point (labeled +), an anti-ideal point (labeled -), or a saddle point (labeled o);

The objects whose position on the map is determined by their coordinates;

The descriptors which correspond to the representation axes with which they are associated (when a PCA precedes the PREFMAP, a biplot from the PCA is studied to interpret the position of the objects as a function of the objective criteria).

The PREFMAP, with the interpretation given by the preference map is an aid to interpretation and decision-making which is potentially very powerful since it allows preference data to be linked to objective data. However, the models associated with the assessors must be adjusted correctly in order that the interpretation is reliable.

Preference scores

The preference score for each object for a given assessor, whose value is between 0 (minimum) and 1 (maximum), is calculated from the prediction of the model for the assessor. The more the product is preferred, the higher the score. A preference order of objects is deducted from the preference scores for each of the assessors.

Contour plot

The contour plot shows the regions corresponding to the various preference consensus levels on a chart whose axes are the same as the preference map. At each point on the chart, the percentage of assessors for whom the preference calculated from the model is greater than their mean preference is calculated. In the regions with cold colors (blue), a low proportion of models give high preferences. On the other hand, the regions with hot colors (red) indicate a high proportion of models with high preferences.

Potential ideal points and admission zone

The potential ideal points and admission zone play a crucial role in guiding research and development teams. Potential ideal points represent the maximum number of cases for which the model shows a preference above the chosen criterion.

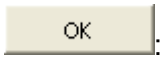
To calculate potential ideal points, XLSTAT uses the maximum preference score, multiplied by a tolerance. This tolerance establishes a threshold beyond which a point is considered potentially ideal. If the number of ideal points exceeds a defined maximum, a k-means analysis is initiated

with k ranging from 2 to this maximum number. Subsequently, the elbow method is employed to determine the optimal number of points.

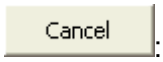
The admission zone, on the other hand, defines a region outside of which finding an ideal point is less likely. It is defined by the minimum and maximum values of each dimension composed of all ideal points (before the k-means analysis, if applicable).

Dialog box

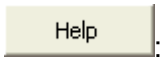
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Preference data: Select the preference data. The table contains the various objects (products) studied in the rows and the assessors in the columns. This is reversed in transposed mode. If column headers have been selected, check that the "Variable labels" option has been activated.

Note: XLSTAT considers that the preferences are the increasing data (the more an assessor likes an object, the higher the preference).

Center: Activate this option is you want to center the preference data before starting the calculations.

Reduce: Activate this option is you want to reduce the preference data before starting the calculations.

X / Configuration: Select the data corresponding to the objective descriptors or to a 2- or 3-dimensional configuration if a method has already been used to generate the configuration. If column headers have been selected, check that the "Variable labels" option has been activated.

Preliminary transformation: Activate this option if you want to transform the data.

- **Normalization**: Activate this option to standardize the data for the X-configuration before carrying out the PREFMAP.
- **PCA (Pearson)**: Activate this option for XLSTAT to transform the selected descriptors using a normalized Principle Components Analysis (PCA). The number of factors used afterwards used for the calculations is determined by the number of **dimensions** chosen.
- **PCA (Covariance)**: Activate this option for XLSTAT to transform the selected descriptors using a non-normalized Principle Components Analysis (PCA). The number of components used afterwards for the calculations is determined by the number of **dimensions** chosen.
- **PLS(Std) / PLS**: Activate this option for XLSTAT to transform the selected descriptors using the components extracted by PLS regression. The latter present the advantage of taking into account covariance structure among Xs as well as among and Ys and Xs and Ys. The (Std) option allows to first standardize the data to remove scale and mean effects. The number of components used afterwards for the calculations is determined by the number of **dimensions** chosen.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the data selected (Y, X, object labels) contains a label.

Objects labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Options tab:

Model: Choose the type of model to use to link the preferences to the X configuration if the option "Find the best model" (see Options tab) has not been activated.

Dimensions: Enter the number of dimensions to use for the PREFMAP model (default value: 2).

Find the best model: Activate this option to allow XLSTAT to find the best model for each assessor.

- **F-ratio**: Activate this option to use the F-ratio test to select the model that is the best compromise between quality of the fit and parsimony in variables. A more complex model is accepted if the p-value corresponding to the F is lower than the significance level.
- **F**: Activate this option to select the model that gives the best p-value based computed the Fisher's F.

Significance level (%): enter the significance level. The p-values of the models are displayed in bold when they are less than this level.

Weights: If you want to weigh the assessors, activate this option, then select the weight corresponding to each observation.

These options are visible only if a PCA based preliminary transformation has been requested.

Supplementary variables: Activate this option if you want to calculate coordinates afterwards for variables which were not used in calculating the factor axes (passive variables as opposed to active variables).

- **Quantitative**: Activate this option if you have supplementary quantitative variables. If column headers were selected for the main table, ensure that a label is also present for the variables in this selection.

Prediction tab:

This tab is not visible if a preliminary PCA transformation was requested.

Prediction: activate this option if you want to select data to use them in prediction mode. If you activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

X / Configuration: Activate this option to select the configuration data to use for the predictions. The first row must not include variable labels.

Object labels: Activate this option if you want to use object labels for the prediction data. The first row must not include variable labels. If this option is not activated, the labels are automatically generated by XLSTAT (PredObs1, PredObs2, etc.).

Variable labels: Activate this option if the first row of the data selections (explanatory variables, observations labels) includes a header.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the correlation matrix for the different variables selected.

Analysis of variance: Activate this option to display the analysis of variance table for the various models.

Model coefficients: Activate this option to display the parameters of the models.

Model predictions: Activate this option to display the predictions of the models.

Preference scores: Activate this option to display preference scores on a scale of 0 to 1.

Ranks of the preference scores: Activate this option to display the ranks for the preference scores.

Sorted objects: Activate this option to display the objects in decreasing order of preference for each of the assessors.

Potential ideal points: Activate this option to display the table of potential ideal points.

- **Tolerance (%)**: Enter the percentage of tolerance to adjust the threshold determining which points are considered potentially ideal. The value must be between 0 and 100: 0 indicates zero tolerance, meaning only points with the maximum preference value are considered ideal. Conversely, 100 implies full tolerance, considering all points as ideal. In other words, a higher tolerance decrease the threshold for a point to be considered ideal. The new threshold is calculated as follows: $$ Threshold = (1 - Tolerance/100) \times Maximum Preference Value $$
- **Maximum number**: Enter the maximum number of ideal points, keeping it within the range of 2 to 10 for a clearer interpretation of the ideal points.

Admission zone: Activate this option to display the admission zone for ideal points. The zone is defined by a minimum and a maximum for each input variable. Out of this zone, it is very unlikely that you find an ideal point.

If a preliminary transformation based on PCA or PLS has been requested, the following options are available:

Eigenvalues: Activate this option to display the eigenvalues corresponding to the different factors extracted (PCA only).

Factor loadings: Activate this option to display the coordinates of the variables (*factor loadings*). The coordinates are equal to the correlations between the components and the initial variables.

Components/Variables correlations: Activate this option to display correlations between the components and the initial variables.

Factor scores: Activate to display the coordinates of the observations (*factor scores*) in the new space created by PCA or PLS. These coordinates are afterwards used for the PREFMAP.

If a preliminary transformation based on PCA has been requested, the following options are available:

Factor loadings: Activate this option to display the coordinates of the variables (*factor loadings*). The coordinates are equal to the correlations between the principal components and the initial variables for normalized PCA.

Components/Variables correlations: Activate this option to display correlations between the principal components and the initial variables.

Factor scores: Activate to display the coordinates of the observations (*factor scores*) in the new space created by PCA. The principal components are afterwards used as explanatory variables in the regression.

Charts (PCA) tab:

This tab is visible only if a PCA based preliminary transformation has been requested.

Correlations charts: Activate this option to display charts showing the correlations between the components and initial variables.

- **Vectors**: Activate this option to display the input variables in the form of vectors.

Observations charts: Activate this option to display charts representing the observations in the new space.

- **Labels**: Activate this option to have observation labels displayed on the charts. The number of labels displayed can be changed using the filtering option.

Biplots: Activate this option to display charts representing the observations and variables simultaneously in the new space.

- **Vectors**: Activate this option to display the initial variables in the form of vectors.

- **Labels**: Activate this option to have observation labels displayed on the biplots. The number of labels displayed can be changed using the filtering option.

Type of biplot: Choose the type of biplot you want to display. See the description section of the PCA for more details.

- **Correlation biplot**: Activate this option to display correlation biplots.
- **Distance biplot**: Activate this option to display distance biplots.
- **Symmetric biplot**: Activate this option to display symmetric biplots.
- **Coefficient**: Choose the coefficient whose square root is to be multiplied by the coordinates of the variables. This coefficient lets you to adjust the position of the variable points in the biplot in order to make it more readable. If set to other than 1, the length of the variable vectors can no longer be interpreted as standard deviation (correlation biplot) or contribution (distance biplot).

Colored labels: Activate this option to show variable and observation labels in the same color as the corresponding points. If this option is not activated the labels are displayed in black color.

Charts tab:

Preference map: Activate this option to display the preference map.

- **Display ideal points**: Activate this option to display the ideal points.
- **Display anti-ideal points**: Activate this option to display the anti-ideal points.
- **Display saddle points**: Activate this option to display the saddle points.
- **Domain restriction**: Activate this option to only display the solution points (ideal, anti-ideal, saddle) if they are within a domain to be defined. Then enter the size of the area to be used for the display: this is expressed as a %age of the area delimited by the X configuration (value between 100 and 500).
- **Vectors length**: The options below are used to determine the lengths of the vectors on the preference map when a vector model is used.
    - **Coefficients**: Choose this option so that the length of the vectors is only determined by the coefficients of the vector model.
    - **R^2**: Choose this option so that the length of the vectors is only determined by the $R^{2\wedge}$ value of the model. Thus the better the model is adjusted, the longer is the corresponding vector on the map.
    - **=**: Activate this option to display the vectors with an equal length.

- **Lengthening factor**: Use this option to multiply the length of all vectors by an arbitrary value (default value: 1)

Circular model:

- **Display circles**: Enter the number of isopreference circles to be displayed.

Contour plot: Activate this option to display the contour plot (see description). Afterwards, you need to choose which criterion is used to determine the % of assessors that prefer products at a given point of the preference map.

- **Threshold / Mean (%)**: Enter the level in % with respect to the preference mean above which an assessor can be considered to like a product (the default value, 100, is the mean).
- **Threshold (value)**: Enter the preference value above which an assessor can be considered to like a product (the default value, 100, is the mean).
- **Color scale**: Choose your colors.

PREFMAP & Contour plot: Activate this option to display the superposition of the preference map and of the contour plot.

- **Number of points**: Three quality levels (64000, 81000, and 100000) are possible. If you notice some defects in the map, you can increase the number of points.

Standardized preferences: Activate this option to standardize preference map values. This option is useful when interpretation of the preference map is less intuitive.

Results

Summary statistics: This table shows the number of non-missing values, the mean and the standard deviation (unbiased) for all assessors and all dimensions of the X configuration (before transformation if that has been requested).

Correlation matrix: This table is displayed to give you a view of the correlations between the various variables selected.

Model selection: This table shows which model was used for each assessor. If the model is not a vector model, the solution point type is displayed (ideal, anti-ideal, saddle) with its coordinates.

Analysis of variance: This table shows the statistics used to evaluate the goodness of fit of the model (R^2, F, and Pr>F). When the p-value (Pr>F) is less than the chosen significance level, it is displayed in bold. If the F-ratio test was chosen in the "Options" tab, the results of the F-ratio test are displayed if it was successful at least once.

Model coefficients: This table displays the various coefficients of the chosen model for each assessor.

Model predictions: This table shows the preferences estimated by the model for each assessor and each product. Note: if the preferences have been standardized, these results therefore apply to the standardized preferences.

Preference scores from 0 to 1: This table shows the predictions on a scale of 0 to 1.

Ranks of the preference scores: This table displays the ranks of the preference scores. The higher the rank, the higher the preference.

Objects sorted by increasing preference order: This table shows the list of objects in increasing order of preference, for each assessor. In other words, the last line corresponds to objects preferred by the assessors according to the preference models.

The **preference map** and the **contour plot** are then displayed. On the preference map, the ideal points are shown by (+), the anti-ideal points by (-) and saddle points by (o).

If the option is enabled and you use Excel 2003 or higher, you can view the superposition of the preference map and contour plot. This chart can be resized, but so that the overlay is maintained after resizing, you must click in the Excel sheet and then again on the graph.

Finally, the **Potential ideal points** and the **admission zone** table and chart are displayed. (see description).

Example

A example of Preference Mapping is available on XLSTAT Help Center:

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

References

Danzart, M., Sieffermann, J. M., & Delarue, J. (2004). New developments in preference mapping techniques: Finding out a consumer optimal product, its sensory profile and the key sensory attributes. In *7th Sensometrics Conference. Davis, CA*.

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

Internal Preference Mapping

Use Internal Preference Mapping (IPM) to analyze the ratings given on P products by J assessors (consumers, experts, …): While External Preference Mapping allows to relate consumers ratings to sensory data (chemical measurements, ratings by experts), for Internal Preference Mapping only preference data is necessary. IPM is based on PCA and adds two options to improve the visualization on the results.

In this section:

Description

Internal Preference Mapping (IPM) is based on Principle Component Analysis (PCA) to allow identifying which products correspond to groups of consumers.

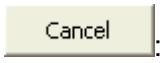
For more information on PCA, you can read the description available in the chapter dedicated to that method. While PCA does not filter out variables, this tool allows removing (after the PCA step) from the plots the assessors that are not well enough displayed on a given 2 dimensional map. The measure of how well a point is projected from a d-dimensional space to a 2-dimensional map is named communality. It can also be understood as the sum of the squared cosines between the vector and the axes of the sub-space.

The biplot that is then produced is not a true biplot as all the retained assessors are moved on a virtual circle surrounding the product points in order to facilitate the visual interpretation.

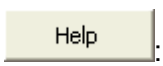
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Products\Assessors table: Select the quantitative data corresponding to P products described by J assessors. If column headers have been selected, check that the "Variable labels" option has been activated.

PCA type: Choose between correlation (normalized PCA), covariance (non normalized PCA) and Spearman to perform PCA on a Spearman correlation matrix.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (products\assessors table, weights, products labels) includes a header.

Product labels: Activate this option if product labels are available. Then select the corresponding data. If the " Assessor labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Filter factors: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %**: Activate this option then enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Activate this option to set the number of factors to take into account.

Standardisation: If the format of your data is "observations/variables", you can choose how correlation (or covariance) will be computed: with denominator (n) or (n – 1).

Rotation: Activate this option if you want to apply a rotation to the factor coordinate matrix.

- **Number of factors**: Enter the number of factors the rotation is to be applied to.
- **Method**: Choose the rotation method to be used. For certain methods a parameter must be entered (Kappa for Orthomax, Tau for Oblimin, and the power for Promax).
- **Kaiser normalization**: Activate this option to apply Kaiser normalization during the rotation calculation.

Supplementary data tab:

Supplementary observations: Activate this option if you want to calculate and represent the coordinates of additional observations. These observations are not taken into account for the computation of the correlation matrix and for the subsequent calculations (we talk of passive observations as opposed to active observations). If the first row of the data selection for supplementary observations includes a header you must activate the "Variable labels for supp. obs" option.

Supplementary variables: Activate this option if you want to calculate coordinates afterwards for variables which were not used in calculating the factor axes (passive variables as opposed to active variables).

- **Quantitative**: Activate this option if you have supplementary quantitative variables. If column headers were selected for the main table, ensure that a label is also present for the variables in this selection.
- **Qualitative**: Activate this option if you have supplementary qualitative variables. If column headers were selected for the main table, ensure that a label is also present for the variables in this selection.
- **Color observations**: Activate this option so that the observations are displayed in different colors depending on the value of the first qualitative variable.
- **Display the centroids**: Activate this option to display the centroids that correspond to the categories of the supplementary qualitative variables.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Pairwise deletion: Activate this option to remove observations with missing data only when the variables involved in the calculations have missing data. For example, when calculating the

correlation between two variables, an observation will only be ignored if the data corresponding to one of the two variables is missing.

Estimate missing data: Activate this option to estimate the missing data before the calculation starts.

- **Mean or mode**: Activate this option to estimate the missing data by using the mean (quantitative variables) or the mode (qualitative variables) for the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data for an observation by searching for the nearest neighbor to the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the correlation or covariance matrix depending on the type of options chosen in the "General" tab.

- **Test significance**: Where a correlation was chosen in the "General" tab in the dialog box, activate this option to test the significance of the correlations.
- **Bartlett's sphericity test**: Activate this option to perform the Bartlett sphericity test.
- **Significance level (%)**: Enter the significance level for the above tests.
- **Kaiser-Meyer-Olkin**: Activate this option to compute the Kaiser-Meyer-Olkin Measure of Sampling Adequacy.

Eigenvalues: Activate this option to display the table and chart (scree plot) of eigenvalues.

Factor loadings: Activate this option to display the coordinates of the variables in the factor space.

Variables/Factors correlations: Activate this option to display correlations between factors and variables.

Factor scores: Activate to display the coordinates of the observations (factor scores) in the new space created by PCA.

Contributions: Activate this option to display the contribution tables for the variables and observations.

Squared cosines: Activate this option to display the tables of squared cosines for the variables and observations.

Filter out assessors: Activate this option if you want to remove on the maps, the assessors for which the communality is below a given threshold.

Charts tab:

Correlations charts: Activate this option to display charts showing the correlations between the components and initial variables.

- **Vectors**: Activate this option to display the initial variables in the form of vectors.

Observations charts: Activate this option to display charts representing the observations in the new space.

- **Labels**: Activate this option to have observation labels displayed on the charts. The number of labels displayed can be changed using the filtering option.

Biplots: Activate this option to display charts representing the observations and variables simultaneously in the new space.

- **Vectors**: Activate this option to display the initial variables in the form of vectors.
- **Labels**: Activate this option to have observation labels displayed on the biplots. The number of labels displayed can be changed using the filtering option.
- **Move to circle**: Activate this option to move all the points corresponding to the assessors moved to a circle that surrounds the points corresponding to the products.

Colored labels: Activate this option to show labels in the same color as the points.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. This includes the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

Correlation/Covariance matrix: This table shows the data to be used afterwards in the calculations. The type of correlation depends on the option chosen in the "General" tab in the dialog box. For correlations, significant correlations are displayed in bold.

Bartlett's sphericity test: The results of the Bartlett sphericity test are displayed. They are used to confirm or reject the hypothesis according to which the variables do not have significant correlation.

Measure of Sample Adequacy of Kaiser-Meyer-Olkin: this table gives the value of the KMO measure for each individual variable and the overall KMO measure. The KMO measure ranges between 0 and 1. A low value corresponds to the case where it is not possible to extract synthetic factors (or latent variables). In other words, observations do not bring out the model that one could imagine (the sample is "inadequate"). Kaiser (1974) recommends not to accept a factor model if the KMO is less than 0.5. If the KMO is between 0.5 and 0.7 then the quality of the sample is mediocre, it is good for a KMO between 0.7 and 0.8, very good between 0.8 and 0.9 and excellent beyond.

Eigenvalues: The eigenvalues and corresponding chart (*scree plot*) are displayed. The number of eigenvalues is equal to the number of non-null eigenvalues.

If the corresponding output options have been activated, XLSTAT afterwards displays the **factor loadings** in the new space, then the correlations between the initial variables and the components in the new space. The **correlations** are equal to the factor loadings in a normalized PCA (on the correlation matrix).

If supplementary variables have been selected, the corresponding coordinates and correlations are displayed at the end of the table.

Contributions: Contributions are an interpretation aid. The variables which had the highest influence in building the axes are those whose contributions are highest.

Squared cosines: As in other factor methods, squared cosine analysis is used to avoid interpretation errors due to projection effects. If the squared cosines associated with the axes used on a chart are low, the position of the observation or the variable in question should not be interpreted.

The **factor scores** in the new space are then displayed. If supplementary data have been selected, these are displayed at the end of the table.

Contributions: This table shows the contributions of the observations in building the principal components.

Squared cosines: This table displays the squared cosines between the observation vectors and the factor axes.

Where a rotation has been requested, the results of the rotation are displayed with the **rotation matrix** first applied to the factor loadings. This is followed by the modified variability percentages associated with each of the axes involved in the rotation. The coordinates, contributions and cosines of the variables and observations after rotation are displayed in the following tables.

Example

A tutorial on how to use Internal Preference Mapping is available on the XLSTAT Help Center:

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

References

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. Monographs on Statistics and Applied Probability, **54**, 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.

Morineau A. and Aluja-Banet T. (1998). Analyse en Composantes Principales. CISIA-CERESTA, Paris.

Liking data analysis

Use this feature to analyze liking data quickly and efficiently.

This function allows you to:

- Determine which products are the most popular
- compare products
- compare assessors or groups of assessors

In this section:

Description

Liking data (also called hedonic data) are among the most collected in sensory analysis. They simply consist in asking the different subjects/consumers/assessors to give a score to the products, generally with a predefined scale on which they have to answer.

Even though the idea behind liking data is very simple, the analysis of these data is rich. The first step is a description of the liking data, with their distribution by product, the differences between sessions, the visualization of the data... A second step, more advanced, consists in performing comparison tests between products and building an internal preference mapping. The last step is based on the study of the agreements between the assessors with the comparison or clustering of groups of assessors or the clustering of the latter.

Structure of the data

There are two different formats:

1. All the assessors data are merged horizontally (horizontal format).
2. All the data are merged vertically (vertical format).

For data entry, XLSTAT asks you to select all the data, and to specify the format type. In the case of vertical format, product and assessor labels are mandatory. Note that if you enter several columns in vertical format, they will be averaged. In the same way, in the case of sessions, these will be studied and then averaged to return to the horizontal format.

Missing values

When a vertical format is entered, XLSTAT will automatically return the data to the horizontal format. This implies the following:

- If several columns are entered and all the values are not missing on a row, then the average is performed on the non-missing values and we don't have any missing value in the horizontal format.
- If an assessor has not seen a product for a session, then the average is performed on the existing sessions. So there will be no missing value in the horizontal format either.
- If an assessor has not seen a product at all, which can be translated by a missing value in the data or simply by the absence of the Product x Assessor combination, then a missing value will be present in the horizontal format (it will be estimated if this option is chosen).

Finally, it should be noted that all results displayed before the end of the data pre-processing (Session description, data Visualization, product and assessor means before centering/scaling) take into account the missing values, unlike the results that follow the data pre-processing.

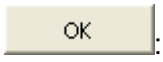
Clustering of the assessors

It is possible to perform a cluster analysis of the assessors. This is done thanks to an Agglomerative Hierarchical Clustering (AHC) based on the Euclidean distance and the Ward criterion. In the case where an automatic choice of the number of clusters is requested, the Hartigan index is used.

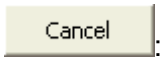
If you want to compare the groups obtained by the clustering of the assessors, you just have to select your preference data in the horizontal format (which are either your initial data or the data displayed by XLSTAT) and the result vector of the groups in the "Groups of assessors" field.

Dialog box

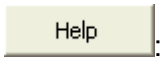
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the

arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Format: Click on horizontal or vertical depending on the structure of your data.

Liking data: Select the data corresponding to the different assessors. If the first row of the selection includes headers, the option "Variable labels" in vertical format or "Assessor labels" in horizontal format must be activated. If you are in vertical format and you select several columns, they will be averaged.

If the format is **horizontal**:

Product labels: Check this option if you want to use the available product labels. If you do not check this option, labels will be created automatically. If a column header has been selected, check that the "Attribute labels" option has been activated.

If the format is **vertical**:

Products: Select the products corresponding to the liking data rows. If a column header has been selected, check that the "Variable labels" option has been activated.

Assessors: Select the assessors corresponding to the liking data rows. If a column header has been selected, check that the "Variable labels" option has been activated.

Sessions: Select the sessions corresponding to the liking data rows. If a column header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Attribute labels (or Variable labels): Activate this option if the first row (or column if in transposed mode) of the selected data (Liking data, Product labels, Assessor labels,Sessions, Groups of assessors) contains a header.

Groups of assessors : Select the data corresponding to the assessor groups. They must have as many values as there are assessors (number of columns in the data).

Options tab:

Center the assessors: Activate this option to center the assessors (mean of each assessor set to 0).

Scale the assessors: Activate this option to scale the assessors (variance of each assessor set to 1).

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals. Default value: 95.

Clustering of the assessors: Activate this option to cluster the assessors (see "Clustering of the assessors" section). Next, decide if you want XLSTAT to **automatically** define a truncation, and therefore the number of clusters to be retained, or if you want to define the **number of clusters** yourself.

Missing data tab :

Details are given in the section "Missing values".

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Product mean**: Activate this option to estimate missing data using the mean of the corresponding product.
- **Assessor mean**: Activate this option to estimate missing data using the mean of the corresponding assessor.

Outputs tab:

Differences between sessions: Activate this option to display the differences between sessions. If you have more than two sessions, the standard deviations between the sessions will be displayed.

Data in horizontal format: Activate this option to display the data in horizontal format.

Product means: Select this option to display the product means table.

Assessor means: Select this option to display the assessor means table.

Tests on product means: Select this option to display the results of product means tests (ANOVA and multiple comparison tests).

Internal preference mapping: Select this option to display tables of results from the internal preference mapping.

Differences between groups: Activate this option to display the results of the tests on the group means product by product (ANOVA and multiple comparison tests).

Interpretation: Activate this option for XLSTAT to calculate an automatic interpretation of the ANOVA results.

Cluster composition: Activate this option to display the cluster composition obtained after truncating the dendrogram.

Charts tab:

Differences between sessions: Activate this option to display the graphs of the differences between the sessions. If you have more than two sessions, the standard deviations between the sessions will be displayed.

Box plots : Activate this option to display the box plot of each product.

Product means: Activate this option to display the graph of the product means.

Assessor means: Activate this option to display the graph of the assessor means.

Visualizing data: Activate this option to display the graph showing the data for the different assessors.

Means charts: Activate this option to display the graphs to view the results of multiple comparisons between products.

Internal Preference Mapping: Activate this option to display the graphs from the internal preference mapping.

Differences between groups: Activate this option to display the graphs showing the results of multiple comparison tests between groups.

Dendrogram: Activate this option to display the dendrogram.

- **Horizontal**: Choose this option to display a horizontal dendrogram.
- **Vertical**: Choose this option to display a vertical dendrogram.
- **Truncated**: Activate this option to display the truncated dendrogram (the dendrogram starts at the level of the truncation).
- **Labels**: Activate this option to display assessor labels (full dendrogram) or clusters (truncated dendrogram) on the dendrogram.
- **Colors**: Activate this option to color each group on the full dendrogram.

Results

Differences between sessions: The table of differences between sessions (standard deviations if you have more than two sessions) is displayed. It allows you to see the possible errors of the assessors or of the data entry (especially if a value is high).

Means of the differences between sessions for each product: The table of the means of the differences between sessions by product is displayed followed by the associated graph. These results can be used to determine if certain products have resulted in session differences.

Means of the differences between sessions for each assessor: The table of the means of the differences between sessions by assessor is displayed followed by the associated graph. These results can be used to determine if any assessors resulted session differences.

Data in horizontal format: Data in horizontal format are displayed without missing values (they are estimated by the option chosen). This data allows you to choose certain options yourself by entering them in an internal preference mapping, an Agglomerative Hierarchical Clustering...

If you have selected groups, the following results will be displayed group by group. In addition, if you have selected the "Center the assessors" option, some results will be given before and after centering the assessors.

Product means: The product means table and the associated bar graph are displayed. This result allows you to determine how much the products are appreciated.

Box plots of the liking scores by product : The box plots of the liking scores for each product are displayed. These allow you to visualize the dispersion of liking data within a product and to compare the dispersions between products.

Visualizing data : A graph allowing to visualize directly the data of the different assessors is displayed. You can choose the assessor to highlight in order to check its data or to compare it to others.

ANOVA: This table allows you to evaluate the explanatory power of the product factor. The explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable (liking data). In other words, if the p-value is significant, we reject the hypothesis that all product means are equal.

Means charts: These graphs allow you to visually compare the means of the products with the associated confidence intervals.

Product/Tukey (HSD): The results of multiple comparison tests of the product means are displayed to determine which products are different from each other and which are similar. The product groups are then given.

Internal preference mapping: The results of the internal preference mapping are displayed. They start with the eigenvalues of the factors as well as the percentages of inertia that each one represents, before displaying the coordinates of the assessors and the coordinates of the products. All these coordinates are also displayed in graphs. Note: if an assessor does not have a representation quality higher than 50% (sum of the squared cosines of the assessor on the axes > 0.5), then it is not displayed.

Differences for each product : The results of the ANOVA, the multiple comparison tests between classes and the associated graphs are displayed for each product.

Clustering of the assessors : The results of the cluster analysis of the assessors are displayed. They contain the obtained dendrogram (truncated if said option has been checked), and the assessor clusters built by truncating the dendrogram.

Example

A tutorial on how to use Liking data analysis is available on the XLSTAT Help Center:

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

References

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.

Panel analysis

Use this tool to check whether your sensory or consumer panel allows to differentiate a series of products. If it does, measure to what extent and make sure that the ratings given by the assessors are reliable.

In this section:

Description

This tool enables chaining different analyzes proposed by XLSTAT, to assess the ability of a panel of J consumers, experts, judges, or assessors (the term assessor is used in the XLSTAT interface), to differentiate P products using K descriptors (variables in the statistical sense) and to control if the ratings are reliable (if repeated measurements are available through different evaluation sessions). A clustering can also be done to identify homogeneous groups among the assessors.

The **first step** consists of a series of ANOVA with the aim to verify for each descriptor if there is a product effect or not. For each descriptor, the table of Type III SS of the ANOVA is displayed for the selected model. Then, a summary table allows comparing the p-values of the product effect for the different descriptors. If you checked the Filter box, the analysis that follows will only be conducted for the descriptors that allow discriminating the products. Different ANOVA models are possible depending on the presence or absence of sessions (repetitions), the willingness to take into account interactions and whether one wants to consider the effect of assessors and sessions as fixed or random.

CAP Table

The CAP (Control of Assessor Performances) table has 2 parts. The left part is a summary of the descriptors. They are sorted according to their product discrimination. If the p-value is less than 0.1, the color will be yellow. If it is less than 0.05, the color will be green. Otherwise, the color will be red. It is exactly the opposite for $product * assessor$ interaction since it is not a positive thing to have a significant interaction. The average of the attribute and the square root of the error end this left part. The right-hand side of the table refers to assessors. Warning, if a filter has been applied, this part will be displayed under the previous one. The assessors are sorted according to their average rank average of the effects produced individually on all descriptors. For a given descriptor, if an assessor does not discriminate the products, he or she will then have a "=". If he/she discriminates the products, and if he/she agrees with the panel (test on his contribution to the assessor*product interaction) , he/she will have a "+", otherwise

he/she will have a "-". Finally, if the assessor has a session effect for the corresponding descriptor (drift-mood), or if he/she is significantly less reliable than other judges from one session to another, he/she is considered non-repeatable and will then have a "!"added.

The **second step** consists of a graphical analysis. For each of the k descriptors that are kept after the ANOVAs, box plots and strip plots are displayed. We can thus see how, for each descriptor, different assessors use the rating scale to evaluate the different products.

The **third step** starts with restructuring the data table, in order to obtain a table containing one row per product and one column per pair of assessor and descriptor (if there are several sessions, then the table contains averages) followed by a PCA (normalized) on this same table. The number of products P is generally less than the product $k \times J$, so we should have at most P factorial axes. We then display as many PCA correlations plots as the number of descriptors, in order to highlight on each plot the points corresponding to the assessors ratings for a given descriptor. This allows to check in one step the extent to which assessors agree or not for each of the k descriptors, once the effect of position and scale is removed (because the PCA is normalized), and to what extent the descriptors are linked or not. To study more precisely the relationship between descriptors, an MFA (Multiple Factor Analysis) plot is displayed.

During the **fourth step** an ANOVA is performed per assessor, and for each of the k descriptors in order to check whether there is a product effect or not. This allows to assess if each assessor is able to distinguish the products using the available descriptors. See the section description of ANOVA method for more details. A summary table is then used to count the number of descriptors for which each assessor was able to differentiate the products. The corresponding percentage is displayed. This percentage is a simple measure of the discriminating power of assessors.

For the **fifth step**, a global table initially presents ratings (averaged over the repetitions if available) for each assessor in rows, and each pair (product, descriptor) in columns. It is followed by a series of P tables and charts to compare, product by product, assessors (averaged over the possible repetitions) for the set of descriptors. These charts can be used to identify strong trends and possible atypical ratings for some assessors.

The **sixth step** allows identifying atypical assessors through the measure for each product of the Euclidean distance of each assessor to an average for all assessors in the space of the k descriptors. A table showing these distances for each product and the minimum and maximum computed over all assessors, allows identifying assessors that are close to or far from the consensus. A chart is displayed to allow visualizing these distances.

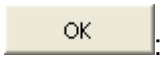
If a "session" variable was selected, the **seventh step** checks if for some assessors there is a session effect, typically an order effect. This is assessed using a Friedman test (or Wilcoxon signed rank test if there are only two sessions). The test is calculated on all products, descriptor by descriptor. Then, for each assessor and each descriptor, we calculate which is the maximum observed range between sessions across products. The product corresponding to the maximum range is indicated on the red triangle. This table is used to identify possible anomalies in the ratings given by some assessors and possibly remove some observations for future analysis.

If for each triple (assessor, product, descriptor) it exists at least one rating, the **eighth step** consists of clustering the assessors. The clustering is first performed on the raw data, then on the standardized data to eliminate possible effects of scale and position.

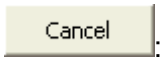
Finally, a pre-formatted table to use the STATIS method is present. This method will allow you to have indications of agreements between assessors and more generally between an assessor and the panel's overall point of view. Moreover, a map of the products will be generated. See the section description of the STATIS method for more details.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.


: Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Y / Descriptors: Select the preference data associated to each descriptor. The table contains the scores given by the assessors for the different descriptors corresponding to a product and to a session. If column headers have been selected, check that the "Variable labels" option has been activated.

Products: Select the data corresponding to the tested products. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

Assessors: Select the data corresponding to the assessors. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

Sessions: Activate this option if more than one tasting session has been organized. Select the data corresponding to the sessions. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the data selected (Attributes, Products, Assessors, Sessions, Observation labels, Observation weights) contains a label.

Observation labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the labels are automatically generated by XLSTAT (Obs1, Obs2…).

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Model: Select the ANOVA model you want to use to identify the non- discriminating descriptors. If the Session option is not active, the two possible models are:

- $Y = Product + Assessor$
- $Y = Product + Assessor + Product * Assessor.$

If the Session option is active, the three possible models are:

- $Y = Product + Assessor + Session$
- $Y = Product + Assessor + Session + Product \times Assessor$
- $Y = Product + Assessor + Session + Product \times Assessor + Product \times Session + Session \times Assessor$

Random effects (Assessor / Session): Activate this option if you want to consider that the Assessor and Repetition effects as well as the interactions involving them are random effects. If this option is not checked, all effects are considered as fixed.

Significance level (%): Enter the significance level that will be used to determine above which level p-values lead to validate the null hypotheses of the various tests that are computed during

the analysis.

Filter descriptors: Activate this option to remove from the analysis all the descriptors for which there is no product effect. You can then specify the threshold p-value above which one can consider there is no product effect. To perform this task, you must not have unchecked "ANOVA Summary" in the Outputs section.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

- **Check each Y separately**: Activate this option to remove observations for each descriptor separately (the sample size will vary from one model to another).
- **For all Y**: Activate this option to remove all observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

ANOVA summaries: Activate this option to display the summaries of the various ANOVA models that are computed.

Assessors' ability to discriminate products: Activate this option to display the tables and charts and allow evaluating the ability of the assessors to differentiate the various products.

Assessor means by (product, descriptor): Activate this option to display the table of the means for each pair (product, descriptor) and, for each product, the table of the means by assessor and descriptor.

Distance to consensus: Activate this option to the table of distances to consensus.

Sessions analysis: Activate this option to assess the reliability of the assessors using the sessions information.

Table for STATIS: Activate this option to display the table formatted to perform a STATIS Analysis.

Charts tab:

Box plots: Activate this option to display the box plots that allow to compare the various assessors for each descriptor.

Strip plots: Activate this option to display the strip plots that allow to compare the various assessors for each descriptor.

PCA plots: Activate this option to display the various plots obtained from the PCA and MFA.

Line plot for each product: Activate this option to display the line plots that allow for each product to compare the assessors for all descriptors.

Line plot of distances to consensus: Activate this option to display the chart that allows to evaluate how far each assessor is from the consensus, product by product.

Dendrogram: Activate this option to display the dendrograms obtained from the clustering of the assessors.

Results

Summary statistics: The tables of descriptive statistics show the simple statistics for all the selected variables. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the descriptors. For qualitative explanatory variables the names of the various categories are displayed along with their respective frequencies.

The **first step** consists of a series of ANOVA with the aim to verify foreach descriptor if there is a product effect or not. For each descriptor, the table of Type III SS of the ANOVA is displayed for the selected model. Then, a summary table allows comparing the p-values of the product effect for the different descriptors.

Then comes the CAP (Control of Assessor Performances) table. See the [description] sections for more details.

The **second step** consists of a graphical analysis. For each of the k descriptors which are kept after the ANOVAs, box plots and strip plots are displayed. We can thus see how, for each descriptor, different assessors use the rating scale to evaluate the different products.

The **third step** starts with restructuring the data table, in order to have a table containing one row per product and one column per pair of assessor and descriptor (if there are several sessions, then the table contains average) followed by a PCA (normalized) on this same table. The number of products P is generally less than the product $k \times J$, so we should have at most P factorial axes. We then display as many PCA correlations plots as the number of descriptors, in order to highlight on each plot the points corresponding to the assessors ratings for a given descriptor. To study more precisely the relationship between descriptors, an MFA (Multiple Factor Analysis) plot is displayed.

During the **fourth step** an ANOVA is performed per assessor, and for each of the k descriptors in order to check whether there is a product effect or not. This allows to assess for each assessor if he/she is able to distinguish the products using the available descriptors. A summary

table is then used to count for each assessor the number of descriptors for which he/she was able to differentiate the products. The corresponding percentage is displayed. This percentage is a simple measure of the discriminating power of assessors.

For the **fifth step**, a global table initially presents ratings (averaged over the repetitions if available) for each assessor in rows, and each pair (product, descriptor) in columns. It is followed by a series of P tables and charts to compare, product by product, assessors (averaged over the possible repetitions) for the set of descriptors. These charts can be used to identify strong trends and possible atypical ratings for some assessors.

The **sixth step** allows identifying atypical assessors through the measure for each product of the Euclidean distance of each assessor to an average for all assessors in the space of the k descriptors. A table showing these distances for each product and the minimum and maximum computed over all assessors, allows identifying assessors that are close to or far from the consensus. A chart is displayed to allow visualizing these distances.

If a "session" variable was selected, the **seventh step** checks if for some assessors there is a session effect, typically an order effect. This is assessed using a Friedman test (or Wilcoxon signed rank test if there are only two sessions). The test is calculated on all products, descriptor by descriptor. Then, for each assessor and each descriptor, we calculate which is the maximum observed range between sessions across products. The product corresponding to the maximum range is indicated on the red triangle.

If for each triple (assessor, product, descriptor) it exists at least one rating, the **eighth step** consists of clustering the assessors. The clustering is first performed on the raw data, then on the standardized data to eliminate possible effects of scale and position.

Finally, a pre-formatted table to use the STATIS method is present.

Example

A tutorial explaining how to use Panel Analysis is available on XLSTAT Help Center:

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

References

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.

Product characterization

Use this tool to identify which descriptors best discriminate a set of products and which characteristics of the products are important in a sensory study.

In this section:

Description

This tool has been developed using the recommendations given by Pr. Jérôme Pagès and Sébastien Lê from the Laboratory for Applied Mathematics at Agrocampus (Rennes, France). It provides the XLSTAT users with a user-friendly tool that helps finding in a sensory study which descriptors are discriminating well a set of products. You can also identify which are the most important characteristics of each product.

All computations are based on the analysis of variance (ANOVA) model. For more details on technical aspects, see the analysis of variance chapter of the XLSTAT help.

The data table must have a given format. Each row should concern a given product, eventually a given session and should gather scores given by an assessor for one or more descriptors associated to the designated product. The dataset must contain the following columns: one identifying the assessor, one identifying the product, eventually one identifying the session, and as many columns as there are descriptors or characteristics.

For each descriptor an ANOVA model is applied to check if the scores given by the assessors are significantly different. The simplest model is:

Score = product effect + assessor effect

If different sessions have been organized (each assessor has evaluated at least twice each product), the session factor can be added and the model becomes:

Score = product effect + assessor effect + session effect

An interaction factor can also be included. We then can test if some combines of the assessors and products are giving higher or lower grades on the descriptors. The model is:

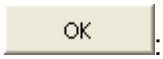
Score = product effect + assessor effect + product effect * assessor effect

The assessor effect is always supposed to be random. It means we consider each assessor to have its own way of giving scores to the products (on the score scale).

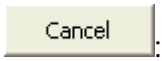
Product characterization is a very efficient tool to characterize products using assessors' preferences.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Descriptors: Select the preference data associated to each descriptor. The table contains the scores given by the assessors for the different descriptors corresponding to a product and to a session. If column headers have been selected, check that the "Variable labels" option has been activated.

Products: Select the data corresponding to the tested products. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

Assessors: Select the data corresponding to the assessors. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

Sessions: Activate this option if more than one tasting session has been organized. Select the data corresponding to the sessions. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the data selected (Y, X, object labels) contains a label.

Observation labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Model: Select the ANOVA model you want to use to identify the non- discriminating descriptors. If the Session option is not active, the two possible models are **Y = Product + Assessor** and **Y= Product + Assessor + Product * Assessor**. If the *Sessions* option is active, the three possible models are **Y = Product + Juge + Session**, **Y = Product + Assessor + Session + Produict * Assessor**, and **Y = Product + Assessor + Session + Product * Assessor + Product * Session + Session * Assessor**.

Sort the adjusted means table: activate this option if you want the adjusted means to be sorted so that similar products and descriptors are close to each other. A principal component analysis is applied to find the best positioning.

Significance level (%): enter the significance level for the confidence intervals.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

- **Check each Y separately**: Activate this option to remove observations for each descriptor separately (the sample size will vary from one model to another).
- **For all Y**: Activate this option to remove all observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Charts tab:

Sensory profiles: Activate this option to display the chart of the sensory profiles.

- **Biplot**: Activate this option to display simultaneously products and Y variables (descriptors).
- **Filter out non discriminating descriptors**: Activate this option to ignore the descriptors that have been identified as non-discriminating in the previous analyses. You can enter the **threshold** above which a descriptor is considered as non discriminating and should be removed.

Results

Summary statistics: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the descriptors. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

Discriminating power by descriptor: This table shows the ordered descriptors from the most discriminating on the products to the least discriminating. Associated V-test and p-values are also displayed.

Model coefficients: This table displays the various coefficients of the chosen model for each combination product-descriptor. Adjusted mean, t test, p-value and confidence interval for each combination are also displayed. Graphics for each product with the coefficients are then displayed.

Adjusted means by product: This table shows the adjusted mean for each combination product-descriptor. The color corresponds to a significant positive effect for the blue color and a significant negative effect for the red color.

Chart with confidence ellipses for the sensory profiles obtained by PCA: this biplot, created following the method described by Husson *et al* (2005) allows to visualize on the same graph the descriptors, as well as the products with a confidence ellipse whose orientation and surface depend on the ratings given by different assessors. These ellipses are calculated using a resampling method. The tables with the coordinates of the products and the corresponding cosines are displayed to avoid misleading interpretations.

Example

An example of product characterization is available on XLSTAT Help Center:

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

References

Husson F. , Lê S. and Pagès J. (2009). SensoMineR dans Evaluation sensorielle - Manuel méthodologique. Lavoisier, SSHA, 3ème édition.

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

Use this tool to analyze the results of a survey run using a JAR (Just About Right) scale, on which the intermediary level 3 corresponds to the preferred value for the consumer.

In this section:

Description

Penalty analysis is a method used in sensory data analysis to identify potential directions for the improvement of products, on the basis of surveys performed on consumers or experts.

Two types of data are used:

Preference data (or liking scores) that correspond to a global satisfaction index for a product (for example, liking scores on a 10 point scale for a chocolate bar), or for a characteristic of a product (for example, the comfort of a car rated from 1 to 10).

Data collected on a JAR (Just About Right) 5 point scale. These correspond to ratings ranging from 1 to 5 (or 1 to 7, or 1 to 9) for one ore more characteristics of the product of interest. In the case of a 5 points JAR scale, 1 corresponds to « Not enough at all », 2 to « Not enough », 3 to « JAR » (*Just About Right*), an ideal for the consumer, 4 to « Too much » and 5 to « Far too much ». For example, for a chocolate bar, one can rate the bitterness, and for the comfort of the car, the sound volume of the engine.

The method, based on multiple comparisons such as those used in ANOVA, consists in identifying, for each characteristic studied on the JAR scale, if the rankings on the JAR scale are related to significantly different results in the liking scores. For example, if a chocolate is too bitter, does that significantly impact the liking scores?

The word penalty comes from the fact that we are looking for the characteristics which can penalize the consumer satisfaction for a given product. The penalty is the difference between the mean of the liking scores for the JAR category, and the mean of the scores for the other categories.

Penalty analysis is subdivided into three phases:

The data of the JAR scale are aggregated: for example, in the case of a 5 points JAR scale, on one hand, categories 1 and 2 are grouped, and on the other hand categories 4 and 5 are

grouped, which leads to a three point scale. We now have three levels: "Not enough", "JAR", and "Too much".

We then compute and compare the means of the liking scores for the three categories, to identify significant differences. The difference between the means of the 2 non-JAR categories and the JAR category is called mean drops.

We compute the penalty and test if it is significantly different from 0.

Penalty Table: Calculate MSE and Standardized Difference

MSE (Mean Squared Error) is a measure of a model's or system's accuracy by comparing predicted values to actual values. The calculation of MSE, used in computing certain values in the penalty table (particularly for calculating the standardized difference), is slightly different from the classic MSE calculation.

Calculating MSE for the JAR Level

$$MSE_{JAR} = \frac{\sum_{i=1}^{n}(y_i - MeanLikingJAR)^2 1_{Score=JAR} + (y_i - MeanLikingNonJAR)^2 1_{Sco}}{n-2}$$

where: - y_i corresponds to the liking value, - $MeanLikingJAR$ corresponds to the mean of preference data with a JAR rating, - $MeanLikingNonJAR$ corresponds to the mean of preference data without a JAR rating.

Calculating the Standardized Difference for the JAR Level

$$StandardizedDifference_{JAR} = \frac{Penalty}{\sqrt{MSE(\frac{1}{n_{JAR}} + \frac{1}{n_{nonJAR}})}}.$$

Calculating MSE for the Too Much and Not Enough Levels

$$MSE_{TooMuch/NotEnough} = \frac{\sum_{i=1}^{n}(y_i - MeanLikingJAR)^2 1_{Score=JAR} + (y_i - MeanLikingT}$$

where: - y_i corresponds to the preference value (liking), - $MeanLikingJAR$ corresponds to the mean of preference data with a JAR rating, - $MeanLikingTooMuch$ corresponds to the mean of preference data with a rating higher than JAR, - $MeanLikingNotEnough$ corresponds to the mean of preference data with a rating lower than JAR.

Calculating the Standardized Difference for the Too Much Level

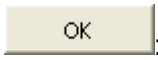
$$StandardizedDifference_{TooMuch} = \frac{EffectOnMean_{TooMuch}}{\sqrt{MSE(\frac{1}{n_{JAR}} + \frac{1}{n_{TooMuch}})}}.$$

Calculating the Standardized Difference for the Not Enough Level

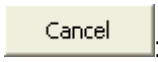
$$StandardizedDifference_{NotEnough} = \frac{EffectOnMean_{NotEnough}}{\sqrt{MSE(\frac{1}{n_{JAR}} + \frac{1}{n_{NotEnough}})}}.$$

Dialog box

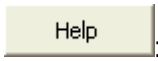
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Liking scores: Select the preference data. Several columns can be selected. If a column header has been selected, check that the "Column labels" option has been activated.

Just about right data: Select the data measured on the JAR scale. Several columns can be selected. If a column header has been selected, check that the "Column labels" option has been activated.

- **Scale**: Select the scale that corresponds to the data (1 -> 5, 1 -> 7, 1 -> 9).

Labels of the 3 JAR levels: Activate this option if you want to use labels for the 3 point JAR scale. There must be three rows and as many columns as in the Just about right data selection. If a column header has been selected, check that the "Column labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selections (Liking scores, Just about right data, labels of the 3 JAR levels) includes a header.

Weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Column labels" option is activated.

Options tab:

Threshold for population size: Enter the % of the total population that should represent a category to be taken into account for multiple comparisons.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to ignore the observations that contain missing data.

Remove by column: Activate this option to remove missing data by column.

Estimate missing data: Activate this option to estimate the missing data by using the mean of the variables.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

Correlations: Activate this option to display the matrix of correlations of the selected dimensions. If all data are ordinal, it is recommended to use the Spearman coefficient of correlation.

3 levels table: Activate this option to display the JAR data once they are collapsed from 5 to 3 categories.

Penalty table: Activate this option to display the table that displays the mean drops for the non-JAR categories, as well as the penalties.

Multiple comparisons: Activate this option to run the multiple comparisons tests on the difference between means. Several methods are available, grouped into two categories: multiple pairwise comparisons, and multiple comparisons with a control, the latter being here the JAR category.

- **Significance level (%)**: Enter the significance level used to determine if the differences are significant or not.

Charts tab:

Stacked bars: Activate this option to display a stacked bars chart that allows visualizing the relative frequencies of the various categories of the JAR scale.

- **3D**: Activate this option to display the stacked bars in three dimensions.

Summary: Activate this option to display the charts that summarize the multiple comparisons of the penalty analysis.

Mean drops vs %: Activate this option do display the chart that displays the mean drops as a function of the corresponding % of the population of testers.

Results

After the display of the basic statistics and the correlation matrix for the liking scores and the JAR data, XLSTAT displays a table that shows for each JAR dimension the frequencies for the 5 levels (or 7 or 9 depending on the selected scale). The corresponding stacked bar diagram is then displayed.

The table of the collapsed data on three levels is then displayed, followed by the corresponding relative frequencies table and the stacked bar diagram.

The penalty table allows to visualize the statistics for the 3 point scale JAR data, including the means, the mean drops, the penalties and the results of the multiple comparisons tests.

Last, the summary charts allow to quickly identify the JAR dimensions for which the differences between the JAR category and the 2 non-JAR categories ("Not enough", "Too much") are significantly different: when the difference is significant, the bars are displayed in red color, whereas they are displayed in green color when the difference is not significant. The bars are displayed in grey when the size of a group is lower than the select threshold (see the Options tab of the dialog box).

The **mean drop vs %** chart displays the mean drops as a function of the corresponding % of the population of testers. The threshold % of the population over which the results are considered significant is displayed with a dotted line.

Example

A tutorial on penalty analysis is available on the XLSTAT Help Center:

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

References

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.

Free Sorting data analysis

Use this function to analyze free sorting data in a quick and efficient way.

This function allows you to:

- to study and visualize the links between products
- to study the agreements between the assessors.

In this section:

Description

Free Sorting tests are becoming more and more popular as part of the sensory characterization of products. They are easy to build and easy to answer. The principle is as follows: each participant is asked to group the products presented to him/her. Each cluster thus represents a set of products that are very similar for the assessor who built it. The number of groups is chosen by the participant. The only prohibited action is to put all products in the same group and to make as many groups as products.

In case the labels of your groups are important, you can use the ACM method which allows you to analyze the labels.

Finally, if you have an incomplete design, you can use missing data. However, only the CA on the co-occurrence matrix method will allow you to analyze your data. Attention, it is strongly recommended that: * Each product is seen the same number of times * Each pair of products is seen the same number of times * The number of subjects is consistent

Methods contained in the a Free Sorting data analysis

The main objective of this analysis is to build a graphical display of the products. For this, three methods are available in XLSTAT:

1. **STATIS**: Data pre-processing is carried out in order to use the STATIS method. Data pre-processing consists of considering each assessor as a complete disjunctive table, where group sizes are then standardized (Llobell, Cariou, Vigneau, Labenne & Qannari, 2020). The STATIS method allows to have indicators of agreement between assessors and to take account of them in the analysis.
2. **CA** on the co-occurrence matrix: a product co-occurrence matrix is build, followed by a Correspondence Analysis (Cariou & Qannari, 2018). Has the advantage of handling

incomplete designs.

3. **MCA**: Multiple Correspondence Analysis is performed on the row data (Van der Kloot & Van Herk, 1991). Has the advantage of analyzing the labels of the groups.

Another objective is to analyze assessors proximities. For this purpose, a co-occurrence matrix between the assessors is build, followed by a Correspondence Analysis (Cariou & Qannari, 2018). Available only if the design is complete.

Structure of the data

Each row represents a product and each column represents a judge. Within each column, you can find the numbers (or names) of the groups to which the product belong. Let's take an example where a judge made a group with products P1 and P3, and a group with products P2 and P4. There will then be a 1 (or G1, "group 1", ...) for P1 and P3 and a 2 (or G2, "group 2", ...) for P2 and P4. If you have an incomplete design, indicate missing values when the subject has not seen the product.

Interpreting the results

The representation of the products (resp. assessors) in the space of k factors allows to visually interpret the proximities between the products (resp. assessors), by means of precautions.

We can consider that the projection of a product or an assessor on a plan is reliable if it is far from the center of the graph.

Number of factors

Two methods are commonly used to determine how many factors must be retained for the interpretation of the results:

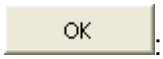
- Watch the decreasing curve of eigenvalues. The number of factors k to be kept corresponds to the first turning point found on the curve.
- We can also use the cumulative variability percentage represented by the factor axes and decide to use only a certain percentage.

Graphic representations

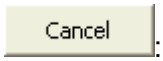
These representations are only reliable if the sum of the variability percentages associated with the axes of the representation space are sufficiently high. If this percentage is high (for example 80%), the representation can be considered as reliable. If the percentage is low, it is recommended to produce representations on several axis pairs in order to validate the interpretation made on the two first factor axes.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.


: Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Free Sorting data: Select the data that correspond to the different assessors. If a column header has been selected, check that the "Assessor labels" option has been activated.

Method: In order to represent the products, 3 methods are available:

- **STATIS**: Activate this option if you want to use the STATIS method after an adapted pre-processing.
- **CA on co-occurrence matrix**: Activate this option if you want to use a Correspondence Analysis on the product co-occurrence matrix.
- **MCA**: Activate this option if you want to use Multiple Correspondence Analysis on raw data.

Product labels: Activate this option if you want to use the available product labels. If you do not activate this option, labels will be created automatically. If a column header has been selected, check that the "Assessor labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Assessor labels: Activate this option if the first row (or column if in transposed mode) of the selected data (Free Sorting data, Product labels) contains a header.

Options tab:

Filter factors: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %**: Enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Set the number of factors to take into account.

Outputs tab:

Genéral: tab:

These outputs concern all the methods of analysis:

Descriptive statistics: Activate this option to display descriptive statistics for all selected assessors.

Assessors analysis: Activate this option if you want to perform a assessors analysis.

- **Co-occurrence matrix**: Activate this option to display the assessors co-occurrence matrix.
- **CA eigenvalues**: Activate this option to display the table of eigenvalues of the CA on the assessors co-occurrence matrix.
- **Assessor coordinates**: Activate this option to display the coordinates of the assessors in the factors space.

STATIS tab:

These outputs only concern the STATIS analysis, and are displayed if this the method has been chosen by the user:

Eigenvalues: Activate this option to display the table of eigenvalues.

Consensus coordinates: Activate this option to display the coordinates of the consensus in the factors space.

RV matrix: Activate this option to display the RV matrix.

Scaling factors: Activate this option to display the scaling factors.

Weights: Activate this option to display the weights created and used by STATIS.

Consensus configuration: Activate this option to display the consensus configuration created by STATIS.

Homogeneity: Activate this option to display homogeneity of the assessors.

RV assessors/consensus: Activate this option to display the RV coefficient between each assessor and the consensus.

Global error: Activate this option to display the error of the STATIS criterion.

Residuals per assessor: Activate this option to display the error of the STATIS criterion for each assessor.

Residuals per product: Activate this option to display the error of the STATIS criterion for each product.

CA tab:

These outputs only concern the CA on the product co-occurrence matrix, and are displayed if this the method has been chosen by the user:

Co-occurrence matrix: Activate this option to display the product co-occurrence matrix.

CA eigenvalues: Activate this option to display the table of eigenvalues of the CA on the co-occurrence matrix.

Product coordinates: Activate this option to display the coordinates of the products in the factors space.

MCA tab:

These outputs only concern the MCA, and are displayed if this the method has been chosen by the user:

Eigenvalues: Activate this option to display the table of eigenvalues of the MCA.

Product coordinates: Activate this option to display the coordinates of the products in the factors space.

Product contributions: Activate this option to display the contributions of the products.

Labels: Activate this option to display the coordinates of the groups in the factor space.

Charts tab:

Genéral: tab:

These outputs concern all analyses:

Display charts on two axes: Activate this option so that XLSTAT does not prompt you to select the axes, and automatically displays the graphs on the first two axes.

Assessor analysis:

- **CA eigenvalues**: Activate this option to display the *scree plot* of the CA eigenvalues.
- **Assessor coordinates**: Activate this option to display the plot of the assessor coordinates in the factors space.

STATIS tab:

These outputs only concern the STATIS analysis, and are only available if this is the method you have chosen:

Eigenvalues: Activate this option to display the *scree plot* of the eigenvalues.

Consensus coordinates: Activate this option to display the plot of the consensus coordinates in the factors space.

Scaling factors: Activate this option to display the bar chart of the scaling factors.

Weights: Activate this option to display the bar chart of the weights created and used by STATIS.

RV assessors/consensus: Activate this option to display the bar chart of the RV coefficient between each assessor and the consensus.

Residual per assessor: Activate this option to display the bar chart of the error of the STATIS criterion for each assessor.

Residual per product: Activate this option to display the bar chart of the error of the STATIS criterion for each product.

Charts of the projected points: Activate this option to display the graphic representing both the products, and the products of each assessor projected in the factor space.

- **Product labels**: Activate this option to display the product labels on the charts.
- **Projected points labels**: Activate this option to display the labels of the projected points.

CA tab:

These outputs only concern the CA on the product co-occurrence matrix, and are only available if this is the method you have chosen:

CA eigenvalues: Activate this option to display the *scree plot* of the CA eigenvalues.

Product coordinates: Activate this option to display the plot of the coordinates of the products in the factors space.

MCA tab:

These outputs only concern the MCA, and are only available if this is the method you have chosen:

Eigenvalues: Activate this option to display the *scree plot* of the MCA eigenvalues.

Product coordinates: Activate this option to display the plot of the coordinates of the products in the factors space.

Labels: Activate this option to display the graphs of the groups in the factor space. If the product coordinate graphs are also selected, then a biplot will be displayed.

Results

Summary statistics: The summary statistics table presents simple statistics for all selected assessors. The groups of each assessor with their respective sizes and percentages of products in each group are displayed.

STATIS:

These results only concern the STATIS analysis, and are only available if this is the method you have chosen:

Eigenvalues and percentages of inertia: The eigenvalues and corresponding chart (*scree plot*) are displayed.

Consensus coordinates: Consensus coordinates in the factors space are displayed, with the corresponding charts (depending on the number of factors chosen).

RV matrix: The matrix of RV coefficients between all assessors is displayed. The RV index is a coefficient of similarity between two assessors included between 0 and 1. The closer it is to 1, the stronger the similarity. This matrix is used by STATIS to calculate the weights of the assessors.

Scaling factor for each assessor: The scaling factors are displayed with the associated bar chart. These scale factors standardize the number of groups of each assessor. The fewer groups an assessor has made, the greater the scale factor.

Weight of each assessor: The weights calculated by STATIS are displayed, with the associated bar chart. The greater the weight, the more the assessor contributed to the consensus. Knowing that STATIS gives more weight to the closest assessors from a global point of view, a much lower weight than the others will mean that the assessor is atypical.

Consensus configuration: The consensus configuration is displayed. It corresponds to the sum of the pre-processed assessor data weighted by the weights of these assessors.

Homogeneity: The homogeneity of the assessors is displayed. It is a value between $1/m$ (m is the number of assessors) and 1, which increases with the homogeneity of the assessors.

RV index between each assessor and the consensus: The RV coefficients between the assessor and the consensus are displayed, with the associated bar chart. Like the weights of STATIS, these coefficients make it possible to detect atypical assessors. The advantage of these coefficients is that they are between 0 and 1, so they are easier to interpret than the weights.

Global error: The global error of the STATIS criterion is displayed. It corresponds to the sum of all residuals (which can be presented by assessor or product).

Residual per assessor: This table and the corresponding bar chart make it possible to visualize the distribution of the residual per assessor. It is thus possible to identify for which assessors STATIS has been less efficient, or in other words, which assessors stand out the most from the consensus.

Residual per product: This table and the corresponding bar chart make it possible to visualize the distribution of the residual per product. It is thus possible to identify for which products STATIS has been less efficient, or in other words, which products stand out the most from the consensus.

Charts of the projected points: The projected points correspond to the projections of the products of each assessor in the factor space. The representation of the projected points superimposed with those of the objects makes it possible to visualize at the same time the diversity of the information brought by the various assessors for a given product, and to visualize the relative distances from two objects according to the various assessors.

CA on co-occurrence matrix:

These results only concern the CA on the product co-occurrence matrix, and are only available if this is the method you have chosen:

Co-occurrence matrix: The co-occurrence matrix between all products is displayed. This symmetrical matrix shows how many times two products were placed in the same group by the assessors.

Eigenvalues and percentages of inertia: The eigenvalues and corresponding chart (*scree plot*) are displayed.

Product coordinates: The coordinates of the products in the factors space are displayed with the corresponding charts (depending on the number of factors chosen).

Multiple Correspondence Analysis:

These results only concern the MCA, and are only available if this is the method you have chosen:

Eigenvalues and percentages of inertia: The eigenvalues and corresponding chart (*scree plot*) are displayed.

Product coordinates: The coordinates of the products in the factors space are displayed with the corresponding charts (depending on the number of factors chosen).

Product contributions: The contributions of the products are displayed. The contributions are helpful for interpreting the plots. The products that have influenced the most the calculation of the axes are those that have the higher contributions.

Labels: the contributions of the groups of each topic are displayed. A biplot is then displayed.

Assessors analysis:

Co-occurrence matrix: The co-occurrence matrix between all assessors is displayed. This symmetrical matrix shows how many times two judges have both placed two different products in the same group.

Eigenvalues and percentages of inertia: The eigenvalues and corresponding chart (*scree plot*) are displayed.

Assessor coordinates: The coordinates of the assessors in the factors space are displayed with the corresponding charts (depending on the number of factors chosen).

Example

A tutorial on how to use Free Sorting data analysis is available on the XLSTAT Help Center:

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

References

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.

Projective mapping data analysis

Use this function to analyze projective mapping data in a quick and efficient way.

This function allows you to:

- Study and visualize the links between products.
- Study the agreements between the assessors.

In this section:

Description

The projective mapping (or Napping) task is one of the so-called "rapid" tests that are becoming increasingly popular in the context of the sensory characterization of products. You ask each subject to place products on a sheet of paper. The data collected are simply the coordinates of the products on the x-axis and y-axis of the sheet of paper. Each subject brings a table with n rows (one per product) and 2 columns. These data can be analyzed with the STATIS method or with Multiple Factor Analysis (MFA). While both methods have the primary objective of synthesizing information to graphically represent the products, they also allow you to determine relationships between the subjects' answers.

Structure of the data

Each row represents a product and the columns are the x-axis and y-axis coordinates for each subject. The data of the subjects are merged vertically.

Interpreting the results

The representation of the products in the space of k factors allows you to visually interpret the proximities between the products, by means of precautions.

We can consider that the projection of a product on a plane is reliable if it is far from the center of the graph.

Number of factors

Two methods are commonly used to determine how many factors must be retained for the interpretation of the results:

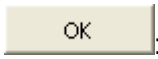
- Watch the decreasing curve of eigenvalues. The number of factors k to be kept corresponds to the first turning point found on the curve.
- We can also use the cumulative variability percentage represented by the factor axes and decide to use only a certain percentage.

Graphic representations

These representations are only reliable if the sum of the variability percentages associated with the axes of the representation space are sufficiently high. If this percentage is high (for example 80%), the representation can be considered reliable. If the percentage is low, a good idea to produce representations on several axis pairs in order to validate the interpretation made on the two first factor axes.

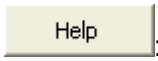
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options, ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.

: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.


: Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Projective mapping data: Select the data that correspond to the different assessors. If a column header has been selected, check that the "Coordinate labels" option has been activated. If the data is **not suitable** for this type of table, it can be **transformed** into a horizontal table, as described above, using the **Create a Products Table.** feature.

Method: In order to represent the products, two methods are available:

- **STATIS**: Activate this option if you want to use the STATIS method.
- **MFA**: Activate this option if you want to use Multiple Factor Analysis.

Product labels: Activate this option if you want to use the available product labels. If you do not activate this option, labels will be created automatically. If a column header has been selected, check that the "Coordinate labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet in the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Coordinate labels: Activate this option if the first row (or column if in transposed mode) of the selected data (Projective mapping data, Product labels, Assessor labels) contains a header.

Assessor labels: Activate this option if you want to use subject labels for the display of results. The number of labels must be the same as the number of subjects in projective mapping data. If the "Coordinate labels" option is activated, the first cell of the selection must include a header. If you do not enable this option, labels will be created automatically.

Options tab:

Filter factors: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %**: Enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Set the number of factors to take into account.

Display charts on two axes: Activate this option so that XLSTAT does not prompt you to select the axes, and automatically displays the graphs on the first two axes.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Estimate missing data: Activate this option to estimate the missing data before the calculation starts.

- **Mean or mode**: Activate this option to estimate the missing data by using the mean (quantitative variables) or the mode (qualitative variables) for the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data for an observation by searching for the nearest neighbour to the observation.

Outputs tab:

STATIS tab:

These outputs only concern the STATIS analysis, and are displayed if this the method has been chosen by the user:

Descriptive statistics: Activate this option to display descriptive statistics for all selected assessors.

Eigenvalues: Activate this option to display the table of eigenvalues.

Consensus coordinates: Activate this option to display the coordinates of the consensus in the factors space.

RV matrix: Activate this option to display the RV matrix.

Scaling factors: Activate this option to display the scaling factors.

Weights: Activate this option to display the weights created and used by STATIS.

Consensus configuration: Activate this option to display the consensus configuration created by STATIS.

Homogeneity: Activate this option to display homogeneity of the assessors.

RV assessors/consensus: Activate this option to display the RV coefficient between each assessor and the consensus.

Global error: Activate this option to display the error of the STATIS criterion.

Residuals per assessor: Activate this option to display the error of the STATIS criterion for each assessor.

Residuals per product: Activate this option to display the error of the STATIS criterion for each product.

MFA tab:

These outputs only concern the MFA, and are displayed if this the method has been chosen by the user:

Descriptive statistics: Activate this option to display descriptive statistics for all selected assessors.

Eigenvalues: Activate this option to display the table of eigenvalues of the MFA.

Product coordinates: Activate this option to display the coordinates of the products in the factors space.

Product contributions: Activate this option to display the contributions of the products.

Squared cosines: Activate this option to display the table of squared cosines of the products.

Lg coefficients: activate this option to display the Lg coefficients of the link between the subjects.

Charts tab:

STATIS tab:

These charts only concern the STATIS analysis, and are only available if this is the method you have chosen:

Eigenvalues: Activate this option to display the *scree plot* of the eigenvalues.

Consensus coordinates: Activate this option to display the plot of the consensus coordinates in the factors space.

Scaling factors: Activate this option to display the bar chart of the scaling factors.

Weights: Activate this option to display the bar chart of the weights created and used by STATIS.

RV assessors/consensus: Activate this option to display the bar chart of the RV coefficient between each assessor and the consensus.

Residual per assessor: Activate this option to display the bar chart of the error of the STATIS criterion for each assessor.

Residual per product: Activate this option to display the bar chart of the error of the STATIS criterion for each product.

Charts of the projected points: Activate this option to display the graphic representing both the products, and the products of each assessor projected in the factor space.

- **Product labels**: Activate this option to display the product labels on the charts.
- **Projected points labels**: Activate this option to display the labels of the projected points.

MFA tab:

These charts only concern the MFA, and are only available if this is the method you have chosen:

Eigenvalues: Activate this option to display the *scree plot* of the MCA eigenvalues.

Product coordinates: Activate this option to display the plot of the coordinates of the products in the factors space.

Charts of the projected points: Activate this option to display the graphic representing both the products, and the products of each assessor projected in the factor space.

- **Product labels**: Activate this option to display the product labels on the charts.
- **Projected points labels**: Activate this option to display the labels of the projected points.

Results

Summary statistics: The summary statistics table presents simple statistics for all selected assessors.

STATIS:

These results only concern the STATIS analysis, and are only available if this is the method you have chosen:

Eigenvalues and percentages of inertia: The eigenvalues and corresponding chart (*scree plot*) are displayed.

Consensus coordinates: Consensus coordinates in the factors space are displayed, with the corresponding charts (depending on the number of factors chosen).

RV matrix: The matrix of RV coefficients between all assessors is displayed. The RV index is a coefficient of similarity between two assessors included between 0 and 1. The closer it is to 1, the stronger the similarity. This matrix is used by STATIS to calculate the weights of the assessors.

Scaling factor for each assessor: The scaling factors are displayed with the associated bar chart. These scaling factors standardize the use of each subject's sheet. The less space a subject has between its products, the larger its scaling factor.

Weight of each assessor: The weights calculated by STATIS are displayed with the associated bar chart. The greater the weight, the more the assessor contributed to the consensus. Knowing that STATIS gives more weight to the closest assessors from a global point of view, a much lower weight than the others will mean that the assessor is atypical.

Consensus configuration: The consensus configuration is displayed. It corresponds to the sum of the pre-processed assessor data weighted by the weights of these assessors.

Homogeneity: The homogeneity of the assessors is displayed. It is a value between $1/m$ (m is the number of assessors) and 1, which increases with the homogeneity of the assessors.

RV index between each assessor and the consensus: The RV coefficients between the assessor and the consensus are displayed, with the associated bar chart. Like the weights of STATIS, these coefficients make it possible to detect atypical assessors. The advantage of these coefficients is that they are between 0 and 1, so they are easier to interpret than the weights.

Global error: The global error of the STATIS criterion is displayed. It corresponds to the sum of all residuals (which can be presented by assessor or product).

Residual per assessor: This table and the corresponding bar chart make it possible to visualize the distribution of the residual per assessor. It is thus possible to identify for which assessors STATIS has been less efficient, or in other words, which assessors stand out the most from the consensus.

Residual per product: This table and the corresponding bar chart make it possible to visualize the distribution of the residual per product. It is thus possible to identify for which products STATIS has been less efficient, or in other words, which products stand out the most from the consensus from one subject to another.

Charts of the projected points: The projected points correspond to the projections of the products of each assessor in the factor space. The representation of the projected points superimposed with those of the objects makes it possible to visualize at the same time the diversity of the information brought by the various assessors for a given product, and to visualize the relative distances from two objects according to the various assessors.

Multiple Factor Analysis:

These results only concern the MFA, and are only available if this is the method you have chosen:

Eigenvalues and percentages of inertia: The eigenvalues and corresponding chart (*scree plot*) are displayed.

Product coordinates: The coordinates of the products in the factors space are displayed with the corresponding charts (depending on the number of factors chosen).

Product contributions: The contributions of the products are displayed. The contributions are helpful for interpreting the plots. The products that have most influenced the calculation of the axes are those that have the higher contributions.

Squared cosines: The projection of a point on an axis, a plane or a 3-dimensional space can be considered reliable if the sum of the cosines squared on the axes of representation is not too far from 1. The cosines squared are displayed in the results proposed by XLSTAT in order to avoid any misinterpretation.

Lg coefficients: the Lg coefficients of the link between the subjects are used to measure how close the subjects are to each other.

Charts of the projected points: The projected points correspond to the projections of the products of each assessor in the factor space. The representation of the projected points superimposed with those of the objects makes it possible to visualize at the same time the diversity of the information brought by the various assessors for a given product, and to visualize the relative distances from two objects according to the various assessors.

Example

A tutorial on how to use projective mapping data analysis is available on the XLSTAT Help Center:

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

References

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.

CATA data analysis

Use this function to analyse CATA (check-all-that-apply) data quickly and efficiently. If the CATA survey includes preference data, this tool can be used to identify drivers of liking or on the opposite, attributes that consumers consider as negative.

In this section:

Description

CATA (check-all-that-apply) surveys have become more and more popular for sensory product characterization since 2007, when it was presented by Adams *et al*. CATA surveys allow to focus on consumers, more representative of the market, instead of trained assessors. They are easy to set up and easy for participants to answer. The principle is that each assessor receives a questionnaire with attributes or descriptors that the respondent may feel, or not, that they apply to one or more products. If it does, he simply needs to check the attribute, otherwise he does not need to do anything. Other questions on different scales may be added to relate the attributes to preferences and liking scores. If participants are asked to give an overall rating to each product of the study, then further analyses and preference modelling is possible. Ares *et al.* (2014) recommend to randomize the order of the CATA questions between assessors to improve the reproducibility

XLSTAT's CATA data analysis tool was developed to automate the analysis of CATA data. Improvements were made by the Addinsoft team in 2020 to better assess the quality of data before analysis.

Let us consider that N assessors were surveyed for P products (one of the products can be a virtual, often ideal, product) on K attributes. The CATA data for the K attributes are assumed to be recorded in a binary format (1 for checked, 0 for not checked). Three formats are currently accepted by XLSTAT:

1. Horizontal format (P x K x N): XLSTAT expects that you have in Excel, a table with P rows, and N groups of K columns all next to each other. You will then only need to specify the value of N, from which XLSTAT will guess K. If you asked each assessor to give his liking, you can add that column within each group of K columns at a position you can let XLSTAT know. In that case each group will have K+1 columns. If one of the products is an ideal product, you can specify its position.

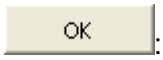
2. Horizontal format (N x K x P): XLSTAT expects that you have in Excel, a table with N rows, and P groups of K columns all next to each other. You will then only need to specify the value of P, from which XLSTAT will guess K. If you asked each assessor to give his liking, you can add that column within each group of K columns at a position you can let XLSTAT know. In that case each group will have K+1 columns. If one of the products is an ideal product, you can specify its position.

3. Vertical format ((N x P) x K): XLSTAT expects that you have in Excel, a table with P x N rows, and K columns. You will then need to select that table. In two additional fields, you need to select the product identifier and the assessor identifier. If you asked each assessor to rate the products, you need to select the column that corresponds to the preference data. If one of the products is an ideal product, you can specify its name so that XLSTAT identifies it.

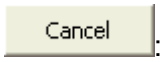
The analyses performed by XLSTAT on CATA data are based on the article by Meyners *et al*. (2013) who investigated in depth the possibilities offered by CATA data.

Dialog box

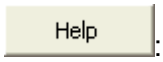
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

CATA data (0/1): Select the CATA data (0/1).

Data format: Choose the format of the data that corresponds to the layout of the CATA data. It can be either **horizontal** or **vertical** (see the description section for further details). If column headers have been selected, check that the "Labels included" option has been activated.

If the format is **horizontal**:

- **(P x K x N)**

Number of assessors: Enter the number of assessors (N). XLSTAT will guess the number of attributes (K).

- **(N x K x P)**

Number of products: Enter the number of products (P). XLSTAT will guess the number of attributes (K).

Position of the ideal product: Choose if the ideal product is at a given position in the CATA table, or if it is at the last position.

Preference data: Choose if the preference (liking) data are at a given position in the CATA table, or if it is at the last position. There must be one preference column for each assessor and one value for each product. It can be missing for the ideal product.

Product labels: Activate this option if Product labels are available. Then select the corresponding data. If the "Labels included" option is activated you need to include a header in the selection.

Assessor labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection.

If the format is **vertical ((N x P) x K)**:

Products: Select the data corresponding to the tested products. Only one column has to be selected. If column headers have been selected, check that the "Labels included" option has been activated.

Assessors: Select the data corresponding to the assessors. Only one column has to be selected. If column headers have been selected, check that the "Labels included" option has been activated.

Preference data: If preference data are available, activate this option and select the data. Only one column has to be selected. If column headers have been selected, check that the "Labels included" option has been activated.

Ideal product: Activate this option if the assessors have qualified an ideal product, and specify how the ideal product is named in the Products field.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if the the data selections include a header.

Options (1) tab:

CATA data validation: Activate this option so that XLSTAT checks the quality of the CATA data.

Cochran's Q test: Activate this option to run a Cochran's Q test.

- **Multiple pairwise comparisons**: Select one of the following methods for multiple pairwise comparisons of products: Critical Difference (Sheskin) procedure or McNemar (Bonferroni) test where the significance level is amended using the Bonferroni approach. For more details on these options please refer to Cochran Test.
- **Differences (or p-values) by attribute**: Activate this option to display for each attribute the table of pairwise differences (critical difference option) or p-values (McNemar Bonferroni option) between products. Bold values represent significant differences.
- **Filter out non significant attributes**: Activate this option to remove the attributes for which the Cochran's Q tests is not significant for a given threshold.

Independence of attributes: Activate this option so that XLSTAT computes a multivariate Chi-square test of the attributes. If the test rejects the null hypothesis of independence, it allows to determine through pairwise comparisons (Fisher's exact tests) which attributes are related.

Correspondence analysis:

Distance: Select the distance to be used for the correspondence analysis (CA): Chi-Square for classical CA, or Hellinger if some terms have low frequencies.

Independence test: Activate this option to run an independence test on the contingency table.

Significance level (%): Enter the significance level for the test. This value is also used to determine when Cochran's Q tests are significant.

Filter factors: You can activate one of the two following options in order to reduce the number of factors displayed:

- **Minimum %**: Activate this option and then enter the minimum percentage that should be reached to determine the number of factors to display.
- **Maximum number**: Activate this option to set the maximum number of factors to take into account when displaying the results.

Options (2) tab:

Filter out products: Activate this option to be able to choose on which products the CATA analysis is performed.

Filter out assessors: Activate this option to be able to choose on which assessors the CATA analysis is performed.

Threshold for population size: Enter the % of the total population that should represent a category to be taken into account for the mean impact analysis within penalty analysis.

If there is an ideal product, this percentage will be the minimum size required of the 2x2 table Ideal/Product.

If there is no Ideal product, the categories are present and absent.

Handling of multiple sessions : in case the assessors have evaluated a product several times, please indicate how XLSTAT should behave.

- **Stop computations** : XLSTAT stops the computations.
- **Merge sessions (priority to 1s)** : XLSTAT aggregates the assessor/product/attribute triplets into one, taking 1 if the attribute has been checked once and 0 otherwise.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Replace missing data by 0: Activate this option if you consider that missing data are equivalent to 0.

Results

CATA data validation

If the corresponding option has been checked in the dialog box, XLSTAT first displays a series of results which allow checking the quality of the CATA data. In order to identify possible anomalies, the data is analyzed first for the assessors and then at the attribute level.

Cochran's Q test

Cochran's Q tests are ran on the Assessors x Products table, independently for each attribute. The first column of the results table gives for each attribute (in rows) the p-values. Pairwise comparisons based on the McNemar-Bonferroni or Marascuilo approach are performed. The next columns give the proportion of assessors that checked the product for the given assessor. The letters in parentheses are only important to be considered if the p-value is significant. They can be used to identify the products responsible of a rejection of the null hypothesis that there is no difference between products. The Cochran's Q test is equivalent to a McNemar test if there are only two products.

Independence of attributes

A multivariate chi-square test is performed to test whether the attributes are independent of each other. If the null hypothesis of independence is rejected, pairwise comparison tests (Fisher's exact tests) are performed to identify which attributes are related.

Correspondence Analysis

CATA data are summarized in a contingency table (sum of the N individual CATA tables - the maximum value for each cell is N). A Correspondance Analysis (CA) is ran to visualize the contingency table. The CA can be based on the chi- square distance or the Hellinger distance (also known as the Bhattacharya distance, which is how it is referred to in the similarity/dissimilarity tool of XLSTAT). The Hellinger distance between two samples depends only on the profiles of these two samples. Hence, the analysis based on the Hellinger distance might be used when the dataset includes terms with low frequency (Meyners *et al.*, 2013).

Attributes with null marginal sum are removed from the correspondence analysis. The following results are displayed: contingency table, test of independence between the rows and the columns, eigenvalues and percentages of inertia, and symmetric or asymmetric plot (respectively for the Chi-square and Hellinger option).

Principal Coordinate Analysis

The tetrachoric correlations (well suited for binary data) between attributes and, when liking scores are available, the biserial correlations (developed to measure the correlation between a binary and a quantitative variable) between liking and attributes are computed and visualized using a Principal Coordinate Analysis (PCOA). The eigenvalues and percentage of inertia and the principal coordinates together with a graphical representation are displayed. The proximities between attributes can be analysed.

Penalty Analysis

If liking scores are available, a penalty analysis is performed. When an ideal product has been evaluated, two analysis are ran, for the must have attributes (P(No)|(Yes) and P(Yes)|(Yes)) and the nice to have attributes (P(Yes)|(No) and P(No)|(No)). In the case where there is no ideal product, these analyses are substituted by a single analysis of presence and absence of the attributes.

A summary table shows the frequencies with which the two situations (P(No)|(Yes) and P(Yes)|(Yes) or P(Yes)|(No) and P(No)|(No) or presence and absence) occurs for each attribute.

The comparison table displays the mean drops in liking between the two situations for each attribute and their significances. This table is illustrated with the mean impact display plot and the mean drops vs % plot. In the case where there is an ideal product the must have and the nice to have analysis are summarized in one mean drops vs % plot.

Attribute analysis

A set of K (one for each attribute) 2x2 tables is displayed, with on the left, the values recorded for the ideal product and at the top, the values obtained for the surveyed products. The table contains the average liking (averaged over the assessors and the products) and the % of all records that correspond to this combination of 0s and/or 1s).

Ideal\Products	**0**	**1**
0	6.2 (12%)	7.4 (8%)
1	5.1 (39%)	7.2 (41%)

For a given attribute,

- If the attribute is checked for the ideal product (second row), then if the preference for the products that are checked (cell [1,1]) is higher than when it is not checked (cell [1,0]), then the attribute is a "**must have** ".
- Symmetrically, if the attribute is not checked for the ideal product (first row), then if the preference for the products that are not checked (cell [0,0]) is higher than when it is not checked (cell [0,1]), then the attribute is a "**must not have** ".

- If (cell [0,1]) > (cell [0,0]) significantly, then the attribute is **nice to have**.
- If the attribute is not checked for the ideal product (first row), and if the preference for the products that are checked (cell [0,1]) is about the same (in XLSTAT we have set this as an absolute difference less than one) as when it is not checked (cell [0,0]), then the attribute is a "**does not harm** ".
- Finally, if the attribute is not a must have and that the preference for the checked products (cell [1,1]) is comparable to the preference fo the unchecked products (cell [1,0]), the attribute **does not influence**.

Some tables could correspond to 3 cases. XLSTAT will only associate each table to one case, but you might want to control the results. XLSTAT will try to relate each 2x2 table to one of the rules defined above in the same order.

In a CATA task containing an ideal product, the objective for the products is to get as close as possible to this ideal product. Therefore, the graph representing the difference in citation (number of checks) between the ideal product and the product in question is very useful. For each attribute, we can see whether the product is similar or different from the ideal product. The more differences an attribute is subject to, the more problematic it is and will be located on the left side of the graph. Conversely, the more for a given attribute the product is similar to the ideal product, the closer the line will be to 0. If the difference is negative, the attribute is not present enough, while if it is positive, it is too present. Finally, the confidence interval is used to determine if the difference with the ideal product is significant.

Example

A tutorial on CATA data analysis is available on the XLSTAT Help Center:

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

References

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.

TCATA data analysis

Use TCATA method (Temporal-Check-All-That-Apply) to analyze your TCATA data. This method allows assessors to continuously select and update the attributes that characterize products as they evolve over time.

In this section:

Description

The TCATA method is a temporal extension of the CATA (Check-All-That-Apply) method developed by Castura *et al* (2016). This method describes the multidimensional sensory properties of products as they evolve over time. Selection and deselection of attributes are tracked continuously over time, allowing assessors to characterize the evolution of sensory changes in products. Graphical results make it possible to visualize the evolution of sensory profiles over time and to compare products. TCATA data must be balanced, meaning that each assessor must evaluate all products in each session. Two different data formats can be used for TCATA data:

- **Binary**: Data has as many rows as there are product product/assessor/attribute and possibly sessions combinations, and as many columns as there are time points. The data is saved in binary form (1 if the attribute is selected, 0 otherwise).
- **Start/End time**: Two columns are expected. For each row corresponding to a product/assessor/attribute combination and possibly session, the "Start time" column contains the time when the attribute was first checked and the "End time" column contains the time when the attribute was unchecked. If the same attribute is selected again later, then a second line with the same product/assessor/attribute combination is required.

Interpreting the results

The main results specific to the TCATA method are as follows:

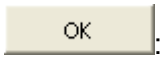
- **Citation proportions curves**: For each product, the citation proportion curves of each attribute are displayed as a function of time. There is an option to smooth the curves. You can also display a reference curve for each attribute. For a given product, this reference curve corresponds to the average proportion of citation for all other products pooled. In order to test the difference between the curve of an attribute and its reference curve a Fisher test or a Khi² test is performed. To avoid overloading the chart, reference curves are displayed only if the difference is significant for a given time. When a significant reference curve is displayed, the corresponding attribute curve is highlighted.

- **Product Differences between products**: A chart is displayed for each product pair. For each attribute we look at whether or not the difference in citation proportions is significant using a Fisher test or Khi² test. When the difference is significant for a given time, the difference in proportion curve is displayed.

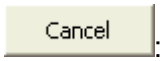
- **Product trajectories**: A correspondence analysis (CA) is performed in order to define the product trajectories. Each row corresponds to a product/time combination and the columns contain the attributes. The factorial coordinates of the rows (product/time) are linked together and thus the product trajectories over time are displayed on the factorial plan. The end point (maximum time) of each trajectory contains the label with the name of the product.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.


: Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Format: Select the format of your TCATA data. The two formats you can use are **Binary** and **Start/End time** (see the description section for further details).

Binary format:

TCATA data (0/1): Select the TCATA binary data. If the "Variable labels" option is activated, the first line of the field will be considered as time values. Otherwise, the vector of time values will be created starting from 0 up to the number of columns in the field.

Start/End time format: **Start time**: Select the data corresponding to the citation start time for each assessor/product/attribute combination. If a column header has been selected, check that the "Variable labels" option has been activated.

End time: Select the data corresponding to the citation end time for each assessor/product/attribute combination. If a column header has been selected, check that the "Variable labels" option has been activated.

Products: Select the data corresponding to the tested products. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

Assessors: Select the data corresponding to the assessors. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

Attributes: Select the data corresponding to the attributes. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

Sessions: Select the data corresponding to the sessions. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the data selections include a header.

Time precision: This option is only available when the "Start/End time" format is used. Choose the level of precision to be used for time data: 1 second or 0.5 seconds.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Replace missing data by 0: Activate this option to replace missing data by 0.

Outputs tab:

CA: Activate this option to perform Correspondence Analysis (CA) and possibly display the product trajectories (see the description section for further details).

- **Eigenvalues**: Activate this option to display the table with eigenvalues of CA.
- **Row coordinates**: Activate this option to display the row coordinates of the table on which the CA was performed, each row corresponds to a product/time combination.
- **Column coordinates**: Activate this option to display the column coordinates of the table on which the AFC was made, each column corresponds to an attribute.

Assessors' agreement: Activate this option to display the table containing assessors' agreement. The closer an assessor agreement is to 1 the more similar his evaluation is to that of the other assessors. The closer his agreement is to 0, the less similar his evaluation is to that of the other assessors.

Assessors' repeatability: If you have selected multiple sessions, activate this option to display the table containing the repeatability of assessors. If the repeatability of an assessor is close to 1, this indicates that he evaluates the same product in the same way between the different sessions. Conversely, if the repeatability is close to 0, the evaluation of the products differs from one session to another.

If you choose to compute CA, you can filter the number of factors to be displayed in the tables of coordinates for rows and columns:

- **Minimum %**: Enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Set the number of factors to take into account.

Charts tab:

Citation proportions:

- **Each attribute by product**: Activate this option to display a bar chart with citation proportions. For each product/attribute combination, a chart will be displayed.
- **Curves by product**: Activate this option to gather citation proportions curves of all attributes on the same chart associated with a product. You can choose to display the attribute reference curves (see the description section for further details).

- **Raw curve**: Activate this option if you want to display raw curves. One point per time slice.
- **Smoothed curve**: Activate this option if you want to display smoothed curves

Difference between products: Activate this option to display for each product pair significant differences in citation proportions (see the description section for further details).

CA chart: If you have chosen in the Options tab to carry out CA, you can choose to display the chart corresponding to the product trajectories (see the description section for further details).

- **Display charts on two axes**: Activate this option if you want to display the chart of product trajectories on the first two axes.
- **Color each product separately**: Activate this option to color each product differently.

If you have chosen to display smoothed curves, different smoothing parameters are offered:

Number of knots:

- **Automatic**: Activate this option to automatically calculate the number as well as the location of the knots (real points) of the smoothed curve. Fewer knots are retained for the portions of the curve having small curvature whereas a larger number of knots is retained for highly curved portions.
- **User defined**: Activate this option to manually set the number of knots in the curve. The coordinates of the latter will be uniformly distributed in abscissa.
- **Tolerance**: Enter the level of smoothing tolerance to be applied to the curves (default value is 0.001).

Test type: The display of reference curves or differences between products requires the calculation of statistical tests (see the description section for further details). The Fisher test is the one used by Castura *et al* (2016), however if you have many products or attributes the large number of tests to perform can greatly slow down the computing time. This is why XLSTAT also proposes to carry out a Khi² test which has the advantage of being much faster.

- **Significance level (%)**: Enter the significance level for the test.

Results

Summary table: A summary table containing the number of sessions, assessors, products and attributes is displayed. Evaluation start and end times are also displayed. If your data is not balanced (see the description section for further details) a table showing the missing combinations is displayed.

Bar chart of citation proportions for each attribute of each product: If you have activated the corresponding option, the bar chart of citation proportions is displayed for each product/attribute combination.

Curves by product: If you have activated the corresponding option, the citation proportions curves of each attribute are displayed for each product.

Significative differences of citation proportions between products: If you have activated the corresponding option, significant differences proportion curves are displayed for each product pair.

Assessors' agreement: If you have activated the corresponding option, the table containing the assessors' agreements is displayed.

Assessors' repeatability: If you have activated the corresponding option, the table containing assessors' repeatability is displayed. **Correspondence Analysis (CA): If you choose to calculate CA, the results will be displayed. The product trajectory chart is also displayed.

Example

An example of TCATA data analysis is available:

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

References

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.

Temporal Dominance of Sensations

Use this tool to analyze Temporal Dominance of Sensations data and identify which descriptors are dominant over time for a set of products.

In this section:

Description

Temporal Dominance of Sensations (TDS) is a temporal multidimensional sensory method (Pineau, Cordelle & Schlich, 2003). Panellists are presented with a list of attributes and asked to choose the dominant ones over consumption of the product. A dominant attribute is the most striking perception at a time, not necessarily the most intense one (Pineau *et al.*, 2009).

The TDS tool of XLSTAT allows visualizing dominant attributes for a set of products.

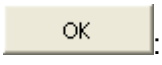
Two data formats are accepted:

1. Dominance format: XLSTAT expects that you have in Excel, a table with as many rows as there are combinations of products, panellists, attributes and potentially sessions, and as many columns as time points. Data must be binary, with 1 if the attribute was selected as dominant at a given time, 0 otherwise. For each combination of panellist*product*session, each attribute must appear only once.

2. Citation format: XLSTAT expects that you have in Excel, a table with one row per attribute selected by a panellist for a given product and potentially given session and one column containing the time of selection. For that format of data, there must be a "start" and a "stop" attribute. For each combination of panellist*product*session, each attribute may appear as many time as it was selected as dominant.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

General tab:

TDS data: Select the TDS data.

Data format: Choose the format of the data that corresponds to the layout of the TDS data. It can be either **dominance (0/1)** or **elicitation** (see the description section for further details). If column headers have been selected, check that the "Variable included" option has been activated.

Products: Select the data corresponding to the tested products. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

Assessors: Select the data corresponding to the assessors. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

Attributes: Select the data corresponding to the attributes. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

Sessions: Select the data corresponding to the sessions. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the the data selections include a header.

Options (1) tab:

Time standardization:

- **None**: Activate this option to keep the data as is (no standardization).
- **Right**: Activate this option to standardize time in order to bring all end times to the same scale. Standardized times are reduced between 0 and 1. for each panelist*product*session times are divided by the maximum time, that is the time of end of evaluation.
- **Left-Right**: Activate this option to standardize time in order to bring all start and end times to the same scale. Standardized times are reduced between 0 and 1. for each panelist*product*session times are reduced by the minimum time, that is the time the first attribute was selected, and divided by the maximum time, that is the time of end of evaluation.

Dominance:

- **Significance level (%)**: Enter the significance level for the test.

Smoothing tolerance

- **Tolerance**: Enter the level of smoothing tolerance to be applied to the DTS curves of the products (default value is 0.001).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Replace missing data by 0: Activate this option if you consider that missing data are equivalent to 0 (only available for dominance type of data).

Remove the observations: Activate this option to remove the observations with missing data.

Charts tab:

TDS curves: Activate this option to display the TDS curves of the products.

- **Smoothing**: Activate this option to smooth curves.
- **Automatic**: Activate this option to automatically calculate the number as well as the location of the knots (real points) of the smoothed curve. Fewer knots are retained for the portions of the curve having small curvature whereas a larger number of knots is retained for highly curved portions (see *Options* section for the appropriate level of smoothing tolerance).
- **User defined**: Activate this option to manually set the number of knots in the curve. The coordinates of the latter will be uniformly distributed in abscissa.
- **Chance limit**: Activate this option to display the chance limit of dominance. The chance limit is the dominance rate that an attribute can obtain by chance and is defined by $P0 = \frac{1}{K}$, K being the number of attributes.

- **Significance limit**: Activate this option to display the significance limit of dominance. The significance limit is the minimum value the dominance rate should equal to be considered as significantly higher than P_0. It is calculated using the confidence interval of a binomial proportionbased on a normal approximation:
- $P_S = P_0 + z_\alpha \sqrt{\frac{P_0(1-P_0)}{J*S}}$ with J the number of panelists and S the number of sessions.

TDS bands: Activate this option to display the TDS bands of the products.

- **Bands (Yes/No)**: Activate this option to display all significantly dominant attributes in one band.
- **Bands by attribute**: Activate this option to display a 2 dimensions band-plot. For each period during which a given attribute is significantly dominant, a band with height relative to the mean dominance rate over the given period is drawn.

Results

Dominance per product per attribute: The tables of dominance per product and attribute show the dominance rate per product and attribute for each time point.

TDS curves: For each product, a plot showing the dominance rate for each attribute plotted against time is displayed. Dominance rates are computed for each time point as the proportion of evaluations (panellist*session) for which the given attribute was assessed as dominant (Pineau *et al.*, 2009). If asked by user, dominance rates are smoothed using cubic spline. If asked by user, the chance limit and significance limit are displayed.

Yes/No bands: For each product, a plot showing the significant dominant attributes as a single band is displayed. The band is composed with stacked colored rectangles (Monterymard *et al.*, 2010). The total height of the band is constant. The x-axis represents time.

Bands by attribute: For each product, a plot showing the significant dominant attributes as bands is displayed. The x-axis represents time and the y-axis shows the different descriptors. The heights of the band are proportional to the mean dominance rates, allowing the user to estimate the importance of each attribute (Galmarini *et al.*, 2016). For each given time period the height of the band is calculated as the mean dominance rate over the period divided by the maximum dominance rate for the given product.

Example

An example of a temporal dominance of sensations (TDS) analysis is available at the XLSTAT Help Center:

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

References

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.

Time-Intensity

Use this tool to analyze Time-Intensity (TI) data and identify the temporal profile of a sensation in a set of products.

In this section:

Description

Time-Intensity is a temporal sensory method first introduced in the 30s. It became notably used in sensory analysis in the 50s (Söjström (1954)) and emerged in the 70s with the improvement of recording instrument.

During TI evaluations, assessors are asked to score the intensity of perception of a single attribute over consumption of the product. Compared to single point measurements, the analysis of the development and decline of particular sensory characteristics may reveal rich information in order to distinguish products or perceptions. This type of analysis may be applicable to a variety of product, ranging from the level of sweetness of a beverage to the feeling left by a lipstick.

TI Data usually consist in several intensity measurements scored by an assessor and recorded at several time steps. Each of those measurements should be associated to a product identifier. XLSTAT also offers the possibility to indicate an assessor as well as a session identifier.

The first step of the TI analysis in XLSTAT is to measure the characteristic parameters on each temporal curve. The initial time of exposure to the stimulus is considered as the first time point on each curve and there are 10 distinct parameters defined as follow:

- I max: peak intensity or maximum observed intensity on the whole curve;
- T start: time point where the reaction to the stimulus is first perceived on the curve, defined as the first intensity value exceeding $X\%$ of the peak intensity;
- T max: time position of the peak intensity on the curve;
- T plateau: time duration around the T max where the measured intensity is greater than $(100 - X)\%$ of the peak intensity;

- T ext: time point of extinction of the perception of the stimulus, defined as the position in time after the peak intensity where the measured intensity is lower than $X\%$ of the peak intensity;
- R increase: slope or rate of intensity increase between T start and T max;
- R decrease: slope or rate of intensity decrease between T max and T ext;
- Area before: the area under the curve before the peak intensity;
- Area after: the area under the curve after the peak intensity;
- Area: the total area under the curve, equal to the sum of Area before and Area after.

Where X is the value of the significance level expressed in $\%$.

The measured curve parameters are displayed in a summarizing table. Time- Intensity curves are expected to match a bell shape pattern. If for some reason, the algorithm detects that one or several curves present pathological characteristics (constant intensity, several maximum, etc…), a message is displayed so that the user can investigate which curve(s) should be removed from the analysis.

The visual control of each curve is an important step in a TI analysis. The user should use its field expertise to make sure curves have meaningful characteristics. To this effect, XLSTAT offers the possibility to display all the recorded curves either on an individual chart or superimposed on a single chart to facilitate the comparison between curves.

In addition to individual time intensity curves, it is also very useful to visually summarize the panel perception of a given stimulus for different products. This can be done easily in XLSTAT by creating a synthetic curve either for the whole data set or for each product identifier. Several techniques are proposed to generate this synthetic curve:

- Average: the synthetic curve is the time step average of all individual curve;
- Parametrized; the synthetic curve is build up from the measured curve parameters;
- Overbosch method: the synthetic curve is created following the approach first proposed in Overbosch (1986);
- Liu and MacFie method: the synthetic curve is created following the approach proposed in Liu (1990).

The last two techniques require that the user specifies an additional parameter which is the desired number of bins for the synthetic curve before and after the peak intensity (the total number of bins of the synthetic curve is therefore twice that number).

Finally an ANOVA is performed on each measured curve parameter separately to assess the product effect and possibly an assessor and or repetition effect. Depending on the selected effects, several model configurations are available to account for potential interactions between products, assessors and sessions. Furthermore, XLSTAT allows the user to treat Assessors and/or Sessions as random effect instead of fixed effect.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

General tab:

Time-Intensity data: Select the Time-Intensity data. Those are the curves as recorded during the evaluation. In the default XLSTAT selection mode (Variable as columns), curves are expected to be organized as rows. Columns then correspond to the successive measurement times. If headers have been selected in the first columns, check that the "Variable labels" option has been activated.

Products: Select the data corresponding to the tested products. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

Assessors: Activate this option if more than one assessor has evaluated the product perception. Select the data corresponding to the assessors. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

Sessions: Activate this option if more than one tasting session has been organized. Select the data corresponding to the sessions. Only one column has to be selected. If column headers have been selected, check that the "Variable labels" option has been activated.

Time: Activate this option if you wish to select a time vector for Time- Intensity data. As for the Time-Intensity Data, the time vector is expected to correspond to one row in the default selection mode. Only one row has to be selected, the time vector is the same for all individual Time-Intensity curve. If the option is not activated, XLSTAT will automatically create a default time vector. If column headers have been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the data selected (Y, X, object labels) contains a label.

Options tab:

Confidence interval (%): Enter the confidence interval that will be used to determine above which level p-values lead to validate the null hypotheses of the various tests that are computed during the analysis.

Model: Select the ANOVA model you want to use. Depending on which effects have been activated on the general tab, several model configurations are proposed to account for possible interactions between products, assessors and sessions.

Random effects (Assessor / Session): Activate this option if you want to consider that the Assessor and Repetition effects as well as the interactions involving them are random effects. If this option is not checked, all effects are considered as fixed.

Create synthetic curve: Activate this option if you want to create a synthetic curve useful to visually summarize the panel perception of a given stimulus for different products (refer to the description for mode detailed information). The several techniques proposed in XLSTAT are:

- **Average**
- **Parametrized**
- **Ovecbosch method**
- **Liu and MacFie method**

One curve per product: Activate this option if you want to create a synthetic curve per product.

Number of bins: if the option create synthetic curve is activated and either the Overbosch method or Liu and MacFie Method has been selected, the user has to enter a number of bins. This number of bins defines the number of data points for the synthetic curve before and after the peak intensity so that the total number of bins of the synthetic curve is twice that number.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

Nearest neighbor: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Curve parameters: Activate this option to display a table with the measured curve parameters.

Synthetic curve table: Activate this option to display a table with the synthetic curve data points.

Analysis of variance: Activate this option to display the summaries of the various ANOVA models that are computed.

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Type III tests of fixed effects: Activate this option to display the type III analysis of variance table.

Product effect: Activate this option to display a summarizing table that shows the product effect for each measured curve parameter.

Interpretation: Activate this option to display additional help on how to understand the displayed results.

Charts tab:

Time-Intensity curves: Activate this option to display the individual Time-Intensity curves on charts.

Display on one chart: Activate this option to display all the curves on a single chart.

Synthetic curve chart: Activate this option to display the created synthetic curves on charts.

Product effect p-values: Activate this option to display a summarizing chart on the product effect for each curve parameter.

Results

Summary statistics: The table of descriptive statistics shows simple statistics for the products, assessors and sessions if selected. The names of the various categories are displayed together

with their respective frequencies.

Measured curve parameters: The table summarizing the measured curve parameters together with the associated products, assessors and sessions identifiers.

Synthetic curves: The table display data points for the created synthetic curves (see description for more detailed information).

Then for each curve parameter a series of ANOVA results is displayed with the aim to verify if there is a product effect or not. For each parameter, the table of Type III SS of the ANOVA is displayed for the selected model (see description for more detailed information). Then, a summary table allows comparing the p-values of the product effect for the different parameters.

Example

A tutorial explaining how to use Time Intensity is available on XLSTAT Help Center:

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

References

Liu Y.H. and MacFie H.J.H. (1990). Methods for averaging time-intensity curves. *Chemical Senses*, **15**, 471-484.

Overbosch P. *et al* . (1986). An improved method for measuring perceived intensity/time relationships in human taste and smell. *Chemical Senses*, **11**, 331-338.

Sensory shelf life analysis

This tool enables you to run sensory shelf life test using assessors' judgments. It is used to find the optimal period for consuming a product using sensorial judgments. XLSTAT-MX uses the parametric survival models to model the shelf life of a product.

In this section:

Description

Sensory shelf life analysis is used to evaluate the ideal period for consumption of a product using sensory evaluation of assessors at different times/dates.

It may happen that the physico-chemical properties of a product are not sufficient to assess the quality of a product with respect to the period in which it is consumed. Frequently, adding sensory evaluation of the product will highlight the best consumption period. In the example of a yogurt, you may have a product that is suitable for consumption but in a sensory evaluation will be too acid or after a certain period will look less attractive.

Methods conventionally used in the analysis of survival data are applicable in this case.

Generally, when conducting this type of sensory tests, the assessors taste the same product at different times/dates. This can be done in different sessions, but it is generally recommended to prepare a protocol that allows to obtain products with different seniority for the test day.

Each assessor will express its opinion on the tested product (like / do not like) and we thus obtain a table of preferences per assessor for each date.

Two formats of input can be used in XLSTAT-MX:

- An assessor $\times$ date table: each column represents a date, each row represents an assessor. There will be two different values ??depending on the preference of the assessor (like / not like).
- A date column and a column with the assessors' names. For each assessor, one enters the date when the change in his preference has been observed. We assume that all judges like the product for the first tasting.

XLSTAT-MX then uses a parametric survival model to estimate a model for the shelf life of the product.

As the exact dates the assessor has change his preference are not known, we use the notion of censoring to set these dates. Thus, if preferences have been collected each week, if an assessor does not like a product after 3 weeks, this assessor is censored by interval between the 2^{nd} and 3^{rd} week. Assume that an assessor appreciates the product all long the study, this assessor is right censored at the last date of the study. Finally, if the assessor likes the product then does not like it and likes it again later in the study, we consider this assessor is left censored at the last date he changed his preference.

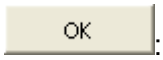
For more details on parametric survival models and censoring, please see the chapter dedicated to these methods. XLSTAT-MX can use an exponential, a Weibull or a log-normal distribution.

As outputs, you will find graphics and parameters of the model.

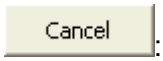
XLSTAT-MX can also add external information to the model using qualitative or quantitative variables associated to each assessor.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Two data format are available (please see the description chapter of this help).

For the "one column per date" option:

Assessor x Date table: Select the table corresponding to the assessors' preference for each date. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Date data: Select the data that correspond to the times/dates that have been recorded. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Positive code: Enter the code used to identify an assessor that appreciated the product. Default value is 1.

Negative code: Enter the code used to identify an assessor that did not appreciate the product. Default value is 0.

For the "one row per assessor" option:

Date data: Select the data that correspond to the times or the dates when it has been observed that an assessor has changed his preference. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Assessor s: Select the data that identify the assessor associated to the event. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Explanatory variables:

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If the variable header has been selected, check that the "Column labels" option has been activated.

Qualitative: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Column labels" option has been activated (see *description*).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selections (time, status and explanatory variables labels) includes a header.

Distribution: Select the distribution to be used to fit your model. XLSTAT-MX offers different distributions including Weibull, exponential, extreme value…

Assessors' labels: In the case of the "one column per date" format, activate this option if you want to select the assessors' names. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Significance level (%): Enter the significance level for the comparison tests (default value 5%). This value is also used to determine the confidence intervals around the estimated statistics.

Initial parameters: Activate this option if you want to take initial parameters into account. If you do not activate this option, the initial parameters are automatically obtained. If a column header has been selected, check that the "Variable labels" option is activated.

Tolerance: Activate this option to prevent the initial regression calculation algorithm taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

Constraints: When qualitative explanatory variables have been selected, you can choose the constraints used on these variables:

$a_1 = 0$: Choose this option so that the parameter of the first category of each factor is set to 0.

$a_n = 0$: Choose this option so that the parameter of the last category of each factor is set to 0.

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the Newton-Raphson algorithm. The calculations are stopped when the maximum number if iterations has been exceeded. Default value: 100.
- **Convergence**: Enter the maximum value of the evolution of the log of the likelihood from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.

Model selection: Activate this option if you want to use one of the two selection methods provided:

- **Forward**: The selection process starts by adding the variable with the largest contribution to the model. If a second variable is such that its entry probability is greater than the **entry threshold value**, then it is added to the model. This process is iterated until no new variable can be entered in the model.
- **Backward**: This method is similar to the previous one but starts from a complete model.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Dates statistics: Activate this option to display statistics for each time/date.

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Model coefficients: Activate this option to display the table of coefficients for the model. The last columns display the hazard ratios and their confidence intervals (the hazard ratio is calculated as the exponential of the estimated coefficient).

Residuals and predictions: Activate this option to display the residuals for all the observations (standardized residuals, Cox-Snell residuals). The value of the estimated cumulative distribution function, the hazard function and the cumulative survival function for each observation are displayed.

Quantiles: Activate this option to display the quantiles for different values of the percentiles (1, 5, 10, 25, 50, 75, 90, 95 and 99 %).

Charts tab:

Preference plot: Activate this option to display the chart corresponding to the number of assessors that likes the product at each date/time.

Preference distribution function: Activate this option to display the charts corresponding to the cumulative preference distribution function (equivalent to the cumulative survival function).

Residuals: Activate this option to display the residual charts.

Results

XLSTAT displays a large number of tables and charts to help in analysing and interpreting the results.

Assessors removed from the analysis: This table displays the assessors that have been removed from the analysis due to a bad coding.

Summary statistics: This table displays descriptive statistics for all the variables selected. For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative variables, the categories with their respective frequencies and percentages are displayed.

Dates statistics: This table displays the number of judges that like the product at each date/time. The associated percentage is also displayed.

Summary of the variables selection: When a selection method has been chosen, XLSTAT displays the selection summary.

Goodness of fit coefficients: This table displays a series of statistics for the independent model (corresponding to the case where there is no impact of covariates, beta=0) and for the

adjusted model.

- **Observations**: The total number of observations taken into;
- **DF**: Degrees of freedom;
- **-2 Log(Like.)**: The logarithm of the likelihood function associated with the model;
- **AIC**: Akaike's Information Criterion;
- **SBC**: Schwarz's Bayesian Criterion;
- **Iterations**: Number of iterations until convergence.

Model parameters: The parameter estimate, corresponding standard deviation, Wald's Khi^2, the corresponding p-value and the confidence interval are displayed for each variable of the model. Confidence intervals are also displayed.

The **residual and predictions** table shows, for each observation, the time variable, the censoring variable, the value of the residuals, the cumulative distribution function, the cumulative survival function and the hazard function..

The **quantiles** associated to the preference curve are presented in a specific table

Charts: Depending on the selected options, charts are displayed. The cumulative preference function or the residuals' plot can be displayed.

Example

A tutorial on how to test sensory shelf life is available on XLSTAT Help Center:

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

References

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.

Generalized Bradley-Terry model

Use this tool to fit a Bradley-Terry model to data obtained from pairwise comparisons.

In this section:

Description

The generalized Bradley-Terry model is used to describe possible outcomes when elements of a set are repeatedly compared with one another in pairs. Consider a set of k elements.

The generalized Bradley-Terry model

For two elements i and j compared in pairs, Bradley and Terry (1952) suggested the following model to evaluate the probability that i is better than j (or i beats j):

$$P(i > j) = \frac{\lambda_i}{\lambda_i + \lambda_j}$$

where λ_i is the skill rating of element i, $\lambda_i \geq 0$.

Several extensions have been proposed for this model. For instance, Agresti (1990) proposed to handle with home-field advantage and Rao and Kupper (1967) developed a model where ties are allowed.

To account for home-field advantage, Agresti (1990) added a parameter δ which measures the strength of this advantage.

$$P(i > j) = \begin{cases} \frac{\delta\lambda_i}{\delta\lambda_i + \lambda_j} \text{ if i at home} \\ \\ \frac{\lambda_i}{\lambda_i + \delta\lambda_j} \text{ if j at home} \end{cases}$$

In the case where ties are allowed between two elements i and j, Rao and Kupper (1967) proposed to include a parameter $\delta > 1$ in the model such that:

$$P(i > j) = \frac{\lambda_i}{\lambda_i + \theta\lambda_j}$$

$$P(i = j) = \frac{\left(\theta^2 - 1\right)\lambda_i\lambda_j}{\left(\lambda_i + \theta\lambda_j\right)\left(\theta\lambda_i + \lambda_j\right)}$$

Inference of model parameters

In the case of a usual Bradley-Terry model, a maximum likelihood estimator of the parameters can be obtained using a simple iterative MM algorithm (Maximization-Minimization, Hunter (2004)). The model (with or without home- field advantage) can be rewritten as a logistic regression model. In this case, a numerical algorithm is used to determine an estimate of the parameters.

In 2012, by considering parameters as random variables, Caron and Doucet proposed a Bayesian approach to overcome the difficulties related to data sparsity. Two methods can be considered:

- Maximizing the log-likelihood by an EM algorithm. For a specific prior distribution on the parameters, this algorithm corresponds to the classical MM algorithm.
- Estimating a posterior distribution of the parameters via a Gibbs sampler.

These two approaches rely on the introduction of latent variables such that the complete likelihood can be written simply. Assume that ω_{ij} is the number of comparisons where i beats j, $\omega_i = \sum_{j=1, j\neq i}^{K} \omega_{ij}$ is the total number of wins of element i and $n_{ij} = \omega_{ij} + \omega_{ji}$ the total number of comparisons between i and j. From the Thurstone interpretation (Diaconis (1988)), the Bradley-Terry model can be written as:

$$P(Y_{ki} < Y_{kj}) = \frac{\lambda_i}{\lambda_i + \lambda_j}$$

where $Y_{ki} \sim \epsilon(\lambda_i)$ and $k \in \{1, \ldots, n_{ij}\}$. To simplify the complete likelihood, a new latent variable Z_{ij} is defined such that:

$$Z_{ij} = \sum_{k=1}^{n_{ij}} min(Y_{kj}, Y_{ki}) \sim \Gamma(n_{ij}, \lambda_i + \lambda_j)$$

In a Bayesian framework, a prior distribution is defined for each parameter. Hence, we assume that the parameters λ_i are distributed according to a Gamma distribution with parameters a and b:

$$P(\lambda) = \prod_{i=1}^{K} \Gamma(\lambda_i; a, b)$$

The prior distribution of the home-field parameter δ is a Gamma $\Gamma(\delta; a_\delta, b_\delta)$ and a flat improper distribution on $[1, +\infty[$ is adopt for the ties parameter θ.

Bayesian EM: This iterative approach aims at maximizing the expected log- likelihood.

Usual model:

At the t-th iteration, the estimate of the parameter λ_i is given by:

$$\lambda_i^{(t)} = \frac{a - 1 + \omega_i}{b + \sum_{j \neq i} \frac{n_{ij}}{\lambda_i^{(t-1)} + \lambda_j^{(t-1)}}}$$

If $a = 1$ and $b = 0$ this estimate corresponds to the MM one.

Model with home-field advantage:

At the t-th iteration, the estimates of the parameters λ_i and δ are:

$$\lambda_i^{(t)} = \frac{a - 1 + \omega_i}{b + \sum_{j \neq i} \frac{\delta^{(t-1)} n_{ij}}{\delta^{(t-1)} \lambda_i^{(t-1)} + \lambda_j^{(t-1)}} + \frac{n_{ji}}{\lambda_i^{(t-1)} + \delta^{(t-1)} \lambda_j^{(t-1)}}}$$

and

$$\delta^{(t)} = \frac{a_\delta - 1 + c}{b_\delta + \sum_{j \neq i} \frac{\lambda_i^{(t)} n_{ij}}{\delta^{(t-1)} \lambda_i^{(t-1)} + \lambda_j^{(t-1)}}}$$

where $c = \sum_{i \neq j} a_{ij}$ and a_{ij} is the number of comparisons where i beats j when i is at home.

Model with ties:

Denote by t_{ij} the number of ties between i and j. At the t-th iteration, the estimates of the parameters λ_i and θ are:

$$\lambda_i^{(t)} = \frac{a - 1 + s_i}{b + \sum_{j \neq i} \frac{s_{ij}}{\lambda_i^{(t-1)} + \Theta^{(t-1)} \lambda_j^{(t-1)}} + \frac{\Theta^{(t-1)} s_{ji}}{\Theta^{(t-1)} \lambda_i^{(t-1)} + \lambda_j^{(t-1)}}}$$

where $s_{ij} = \omega_{ij} + t_{ij}$ and $s_i = \sum_{j \neq i} s_{ij}$. We have:

$$\Theta^{(t)} = \frac{1}{2c^{(t)}} + \sqrt{1 + \frac{1}{4c^{(t)^2}}}$$

and

$$c^{(t)} = \frac{2}{T} \sum_{j \neq i} \frac{s_{ij} \lambda_j^{(t)}}{\lambda_i^{(t-1)} + \Theta^{(t-1)} \lambda_j^{(t-1)}}$$

with $T = \frac{1}{2} \sum_{j \neq i} t_{ij}$ the total number of ties.

Sampling: this approach is based on the Gibbs sampler.

Usual model:

We used the following algorithm to estimate parameter λ_i:

1. For $1 \leq i < j \leq K$ s.t. $n_{ij} > 0$,

$$Z_{ij}^{(t)} | X, \lambda^{(t-1)} \sim \Gamma \left(n_{ij}, \lambda_i^{(t-1)} + \lambda_j^{(t-1)} \right)$$

2. For $1 \leq i \leq K$,

$$\lambda^{(t)} | X, Z^{(t)} \sim \Gamma(a + \omega_i, b + \sum_{i<j | n_{ij}>0} Z_{ij}^{(t)} + \sum_{j<i | n_{ij}>0} Z_{ji}^{(t)})$$

Model with home-field advantage :

We used the following algorithm to estimate the parameters λ_i and δ:

1. For $1 \leq i < j \leq K$ s.t. $n_{ij} > 0$,

$$Z_{ij}^{(t)} | X, \lambda^{(t-1)}, \delta^{(t-1)} \sim \Gamma \left(n_{ij}, \delta^{(t-1)} \lambda_i^{(t-1)} + \lambda_j^{(t-1)} \right)$$

2. For $1 \leq i \leq K$,

$$\lambda^{(t)} | X, Z^{(t)}, \delta^{(t-1)} \sim \Gamma(a + \omega_i, b + \delta^{(t-1)} \sum_{i \neq j | n_{ij}>0} Z_{ij}^{(t)} + \sum_{j \neq i | n_{ij}>0} Z_{ji}^{(t)})$$

Then,

$$\delta^{(t)} | X, Z^{(t)}, \lambda^{(t-1)} \sim \Gamma \left(a_\delta + c, b_\delta + \sum_{i=1}^{K} \lambda_i^{(t)} \sum_{j \neq i | n_{ij}>0} Z_{ij}^{(t)} \right)$$

Model with ties:

We used the following algorithm to estimate the parameters λ_i and θ:

1. For $1 \leq i < j \leq K$ s.t. $n_{ij} > 0$,

$$Z_{ij}^{(t)}|X, \lambda^{(t-1)}, \Theta^{(t-1)} \sim \Gamma\left(s_{ij}, \lambda_i^{(t-1)} + \Theta^{(t-1)}\lambda_j^{(t-1)}\right)$$

2. For $1 \leq i \leq K$,

$$\lambda^{(t)}|X, Z^{(t)}, \Theta^{(t-1)} \sim \Gamma(a + s_i, b + \sum_{i \neq j | s_{ij} > 0} Z_{ij}^{(t)} + \Theta^{(t-1)} \sum_{j \neq i | n_{ij} > 0} Z_{ji}^{(t)})$$

Then,

$$\Theta^{(t)}|X, Z^{(t)}, \lambda^{(t)} \sim P\left(\Theta|X, Z^{(t)}, \lambda^{(t)}\right)$$

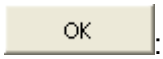
with:

$$P\left(\Theta|X, Z^{(t)}, \lambda^{(t)}\right) \propto \left(\Theta^2 - 1\right)^T exp\left(-\sum_{i \neq j | s_{ij} > 0} Z_{ij}\Theta\right)$$

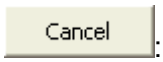
These two methods lead to posterior distributions on the model parameters. However, only the sampling approach allows to estimate the parameters of the complete model.

Dialog box

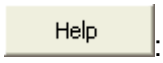
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the

arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Data format: Select the format of the data

- **Two-way table**: Activate this option to select data in a contingency table (wins in rows and losses in column). Only the classical model can be used.
- **Pairs/Variables table**: Activate this option to select data presented in the form of two tables. The pairs table corresponds to the meetings between the elements. The variables table contains the results of each meeting. The first column is the number of wins of the first element and the second column its number of losses. A third optional column can contain the number of ties.

If headers have been selected with the data, make sure the "Labels" option is checked.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels: Activate this option if headers have been selected with the input data.

Options tab:

Inference method: select the inference method

- **Numerical**: The model is rewritten as a logistic regression (see section description). Ties are not allowed.
- **Bayesian EM**: The parameters are supposed to be distributed as a Gamma distribution. The inference is done via an EM algorithm which aims at updating the prior distributions. The parameters of the complete model (with home-field advantage and ties) cannot be inferred with this algorithm.

- **Sampling**: The parameters are supposed to be distributed as a Gamma distribution. The posterior distribution is obtained by a Gibbs sampler.

Options:

- **Home**: Select this option to take home-field advantage into account. In this case, the order of the elements in the pairs table is of importance. The first element is supposed to be at home.
- **Ties**: Select this option if ties are allowed. If the option is enabled, the variables table must have 3 columns. Only the sampling method makes it possible to take into account both home advantage and equality.

Confidence interval (%): Enter the confidence level of the confidence interval of the parameters.

Stop conditions:

- Iterations /Number of simulations: Maximal number of iterations.
- Temps maximum: Maximal allocated time (in second).
- Convergence: Threshold of convergence.

Prior parameter: This option is active only if the inference method is Bayesian EM or Sampling.

- **Scale**: Scale parameter of the Gamma distribution
- **Shape**: Shape parameter of the Gamma distribution.

Outputs tab:

Descriptive statistics: Activate this option to compute and display the statistics that correspond to each element.

Likelihood-based criterion: Activate this option to calculate and to display the likelihood, the BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion).

Probabilities of winning: Activate this option to calculate and to display the probabilities of winning according to model options.

Excel Formulas: Activate this option to display the probability calculation formula when you click on a cell in the probability table.

Charts tab:

Convergence graph: Activate this option to display the evolution of model parameters for the Sampling approach.

Balloon plot: Activate this option to display a graph for quick probability analysis.

Results

Summary statistics: This table displays the descriptive statistics for each element

Estimated parameters: the estimates of the model parameters are given in this table. The standard error and the confidence interval are also provided for each parameter.

Likelihood-based criterion: In this table, several likelihood-based criteria are given (-2*log(Likelihood), BIC, AIC).

Probabilities of winning: This table provides the probability that element i (in row) beats element j (in column), given the model parameters.

Convergence graph: This chart displays for each parameter the evolution of the parameter and the corresponding confidence interval.

Balloon plot: For each of the probability tables, a graph showing the disparities with the size of the circles and the values with the colors is displayed.

Example

An example of the use of Bradley-Terry model is available on XLSTAT Help Center:

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

References

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.

Generalized Procrustes Analysis (GPA)

Use Generalized Procrustes Analysis (GPA) to transform several multidimensional configurations so that they become as much alike as possible. A comparison of transformed configurations can then be carried out.

In this section:

Description

Procrustes (or Procustes), which in ancient Greek means "the one who lengthens while stretching", is a character of the Greek mythology. The name of the gangster Procrustes is associated to the bed that he used to torture the travelers to whom he proposed the lodging. Procrustes installed his future victim on a bed with variable dimensions: short for the tall ones and long for the small ones. According to case's, he cut off with a sword what exceeded out of the bed or stretched the body of the traveler until bringing the size of the traveler to that of the bed, by using a mechanism that Hephaistos had manufactured for him. In both cases the torment was appalling. Theseus, while traveling to Athens, met the robber, discovered the trap and laid down slantwise on the bed. When Procrustes adjusted the body of Theseus, he did not understand the situation immediately and remained perplexed giving Theseus the time to slit with his sword the brigand in two equal parts.

Concept

We define by configuration an n x p matrix that corresponds to the description of n objects (or individuals/cases/products) on p dimensions (or attributes/variables/criteria/descriptors).

We name consensus configuration the mean configuration computed from the m configurations. Procrustes Analysis is an iterative method that allows to reduce, by applying transformations to the configurations (rescaling, translations, rotations, reflections), the distance of the m configurations to the consensus configuration, the latter being updated after each transformation.

Let us take the example of 5 experts rating 4 cheeses according to 3 criteria. The ratings can go from 1 to 10. One can easily consider that an expert tends to be harder in his notation, leading to a shift to the bottom of the ratings, or that another expert tends to give ratings around the average, without daring to use extreme ratings. To work on an average configuration could lead to false interpretations. One can easily see that a translation of the ratings of the first expert is

necessary, or that rescaling the ratings of the second expert would make his ratings possibly closer to those of the other experts.

Once the consensus configuration has been obtained, it is possible to run a PCA (Principal Components Analysis) on the consensus configuration in order to allow an optimal visualization in two or three dimensions.

Structure of the data

There exist two cases:

1. If the number and the designation of the p dimensions are identical for the m configurations, one speaks in sensory analysis about conventional profiles.

2. If the number p and the designation of the dimensions varies from one configuration to the other, one speaks in sensory analysis about free profiles, and the data can then only be represented by a series of m matrices of size n x p(k), k=1,2, …, m.

For the entering of the data, XLSTAT expects an n x (p x m) table, corresponding to m contiguous configurations. If the data is **not suitable** for this type of table, it can be **transformed** into a horizontal table, as described above, using the **Create a Products Table.** feature.

Data transposition

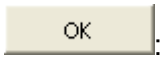
It sometimes occurs that the number (m x p) of columns exceeds the limits of Excel. To get around that drawback, XLSTAT allows you to use transposed tables. To use transposed tables (in that case all tables that you want to select need to be transposed), you only need to click the blue arrow at the bottom left of the dialog box, which then becomes red.

Algorithms

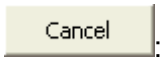
XLSTAT is the unique product offering the choice between the two main available algorithms: the one based on the works initiated by John Gower (1975), and the later one described in the thesis of Jacques Commandeur (1991). Which algorithm performs best (in terms of least squares) depends on the dataset, but the Commandeur algorithm is the only one that allows to take into account missing data; by missing data we mean here that for a given configuration and a given observation or row, the values were not recorded for all the dimensions of the configuration. The later case can happen in sensory data analysis if one of the judges has not evaluated a product.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Configurations: Select the data that correspond to the configurations. If a column header has been selected, check that the "Dimension labels" option has been activated. If the data is **not suitable** for this type of table, it can be **transformed** into a horizontal table, as described above, using the **Create a Products Table.** feature.

Number of configurations: Enter the number of contiguous configurations in the configurations table.

Number of variables per table:

- **Equal**: Choose this option if the number of variables is identical for all the tables. In that case XLSTAT determines automatically the number of variables in each table.
- **User defined**: Choose this option to select a column that contains the number of variables contained in each table. If the "Variable labels" option has been activated, the first row must correspond to a header.

Configuration labels: Check this option if you want to use the available configuration labels. If you do not check this option, labels will be created automatically (C1, C2, etc.). If a column header has been selected, check that the "Dimension labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Dimension labels: Activate this option if the first row (or column if in transposed mode) of the selected data (configurations, configuration labels, object labels) contains a header.

Object labels: Check this option if you want to use the available configuration labels. If you do not check this option, labels will be created automatically (Obs1, Obs 2, etc.). If a column header has been selected, check that the "Dimension labels" option has been activated.

Method: Select the algorithm you want to use:

- **Commandeur**: Activate this option to use the Commandeur algorithm (see the section description section for further details).
- **Gower**: Activate this option to use the Gower algorithm (see the section description section for further details).

Options tab:

Scaling: Activate this option to rescale the matrices during the GPA.

Rotation/Reflect ion: Activate this option to perform the rotation/reflection steps of the GPA.

PCA: Activate this option to run a PCA at the end of the GPA steps.

Filter factors: You can activate one of the following two options in order to reduce the number of factors which are taken into account after the PCA.

- **Minimum %**: Activate this option then enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Activate this option to set the number of factors to take into account.

Tests:

- **Consensus test**: Activate this option to use a permutation test that allows to determine if a consensus is reached after the GPA transformations.
- **Dimensions test**: Activate this option to use a permutation test that allows to determine what is the appropriate number of factors to keep.

Number of permutations: Enter the number of permutations to perform for the tests (Default value: 300)

Significance level (%): Enter the significance level for the tests.

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 100.
- **Convergence**: Enter the maximum value of the evolution in the convergence criterion from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.00001.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Replace missing data by 0: Activate this option to replace missing data by 0.

Remove the observations: Activate this option to remove observations with missing data.

Ignore missing data: Activate this option to use the Ten Berge algorithm to handle missing data.

Outputs tab:

PANOVA table: Activate this option to display the PANOVA table.

Residuals by object: Activate this option to display the residuals for each object.

Residuals by configuration: Activate this option to display the residuals for each configuration.

Scaling factors: Activate this option to display the scaling factors applied to each configuration.

Rotation matrices: Activate this option to display the rotation matrices corresponding to each configuration.

The following options are available only if a PCA has been requested:

Eigenvalues: Activate this option to display the eigenvalues of the PCA.

Consensus configuration: Activate this option to display the coordinates of the dimensions for the consensus configuration.

Configurations: Activate this option to display the coordinates of the dimensions for each configuration.

Objects coordinates: Activate this option to display the coordinates of the objects after the transformations.

- **Presentation by configuration**: Activate this option to display one table of coordinates per configuration.

- **Presentation by object**: Activate this option to display one table of coordinates per object.

Charts (PCA) tab:

The following options are available only if a PCA has been requested:

Eigenvalues: Activate this option to display the scree plot.

Correlations charts: Activate this option to display the correlations charts for the consensus configuration and individual configurations.

- **Vectors**: Activate this option to display the dimensions in the form of vectors.

Objects coordinates: Activate this option to display the maps showing the objects.

- **Presentation by configuration**: Activate this option to display a chart where the color depends on the configuration.
- **Presentation by object**: Activate this option to display a chart where the color depends on the object.

Full biplot: Activate this option to display the biplot showing both the objects and the dimensions of all configurations.

Colored labels: Activate this option to show variable and observation labels in the same color as the corresponding points. If this option is not activated the labels are displayed in black color.

Type of biplot: Choose the type of biplot you want to display. See the description section of the PCA for more details.

- **Correlation biplot**: Activate this option to display correlation biplots.
- **Distance biplot**: Activate this option to display distance biplots.
- **Symmetric biplot**: Activate this option to display symmetric biplots.
- **Coefficient**: Choose the coefficient whose square root is to be multiplied by the coordinates of the variables. This coefficient lets you adjust the position of the variable points in the biplot in order to make it more readable. If set to other than 1, the length of the variable vectors can no longer be interpreted as standard deviation (correlation biplot) or contribution (distance biplot).

Charts tab:

Residuals by object: Activate this option to display the bar chart of the residuals for each object.

Residuals by configuration: Activate this option to display the bar chart of the residuals for each configuration.

Scaling factors: Activate this option to display the bar chart of the scaling factors applied to each configuration.

Test histograms: Activate this option to display the histograms that correspond to the consensus and dimensions tests.

Results

PANOVA table: Inspired from the format of the analysis of variance table of the linear model, this table allows to evaluate the relative contribution of each transformation to the evolution of the variance. In this table are displayed the residual variance before and after the transformations, the contribution to the evolution of the variance of the rescaling, rotation and translation steps. The computing of the Fisher's F statistic allows to compare the relative contributions of the transformations. The corresponding probabilities allow to determine whether the contributions are significant or not.

Residuals by object: This table and the corresponding bar chart allow to visualize the distribution of the residual variance by object. Thus, it is possible to identify for which objects the GPA has been the less efficient, in other words, which objects are the farther from the consensus configuration.

Residuals by configuration: This table and the corresponding bar chart allow to visualize the distribution of the residual variance by configuration. Thus, it is possible to identify for which configurations the GPA has been the less efficient, in other words, which configurations are the farther from the consensus configuration.

Scaling factors for each configuration: This table and the corresponding bar chart allow to compare the scaling factors applied to the configurations. It is used in sensory analysis to understand how the experts use the rating scales.

Rotation matrices: The rotation matrices that have been applied to each configuration are displayed if requested by the user.

Results of the consensus test: This table displays the number of permutations that have been performed, the value of Rc which corresponds to the proportion of the original variance explained by the consensus configuration, and the quantile corresponding to Rc, calculated using the distribution of Rc obtained from the permutations. To evaluate if the GPA is effective, one can set a confidence interval (typically 95%), and if the quantile is beyond the confidence interval, one concludes that the GPA significantly reduced the variance.

Results of the dimensions test: This table displays for each factor retained at the end of the PCA step, the number of permutations that have been performed, the F calculated after the GPA (F is here the ratio of the variance between the objects, on the variance between the configurations), and the quantile corresponding to F calculated using the distribution of F obtained from the permutations. To evaluate if a dimension contributes significantly to the quality of the GPA, one can set a confidence interval (typically 95%), and if the quantile is beyond the confidence interval, one concludes that factor contributes significantly. As an indication are also

displayed, the critical values and the p-value that corresponds to the Fisher's F distribution for the selected alpha significance level. It may be that the conclusions resulting from the Fisher's F distribution is very different from what the permutations test indicates: using Fisher's F distribution requires to assume the normality of the data, which is not necessarily the case.

Results for the consensus configuration:

Objects coordinates before the PCA: This table corresponds to the mean over the configurations of the objects coordinates, after the GPA transformations and before the PCA.

Eigenvalues: If a PCA has been requested, the table of the eigenvalues and the corresponding scree plot are displayed. The percentage of the total variability corresponding to each axis is computed from the eigenvalues.

Correlations of the variables with the factors: These results correspond to the correlations between the variables of the consensus configuration before and after the transformations (GPA and PCA if the latter has been requested). These results are not displayed on the circle of correlations as they are not always interpretable.

Objects coordinates: This table corresponds to the mean over the configurations of the objects coordinates, after the transformations (GPA and PCA if the latter has been requested). These results are displayed on the objects charts.

Results for the configurations after transformations:

Variance by configuration and by dimension: This table allows to visualize how the percentage of total variability corresponding to each axis is divided up for the configurations.

Correlations of the variables with the factors: These results, displayed for all the configurations, correspond to the correlations between the variables of the configurations before and after the transformations (GPA and PCA if the latter has been requested). These results are displayed on the circle of correlations.

Objects coordinates (presentation by configuration): This series of tables corresponds to the objects coordinates for each configuration after the transformations (GPA and PCA if the latter has been requested). These results are displayed on the first series of objects charts.

Objects coordinates (presentation by object): This series of tables corresponds to the objects coordinates for each configuration after the transformations (GPA and PCA if the latter has been requested). These results are displayed on the second series of objects charts.

Example

A tutorial on Generalized Procrustean Analysis is available on XLSTAT Help Center. To view this tutorial go to:

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

References

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.

Multiple Factor Analysis (MFA)

Use the Multiple Factor Analysis (MFA) to simultaneously analyze several tables of variables, and to obtain results, particularly charts, that allow to study the relationship between the observations, the variables and the tables. Within a table, the variables must be of the same type (quantitative table, qualitative table or frequency table), but the tables can be of different types.

In this section:

Description

Multiple Factor Analysis (MFA) makes it possible to analyze several tables of variables simultaneously, and to obtain results, in particular charts, that allow studying the relationship between the observations, the variables and tables (Escofier and Pagès, 1984).

The MFA is a synthesis of the PCA (Principal Component Analysis) for quantitative tables, the MCA (Multiple Correspondence Analysis) for qualitative tables and the CA (Correspondence Analysis) for frequency tables. The methodology of the MFA breaks up into two phases:

We successively carry out for each table a PCA, an MCA or a CA according to the type of the variables of the table. One stores the value of the first eigenvalue of each analysis to then weight the various tables in the second part of the analysis.

One carries out a weighted PCA on the columns of all the tables, knowing that the tables of qualitative variables are transformed into complete disjunctive tables, each indicator variable having a weight that is a function of the frequency of the corresponding category. The weighting of the tables prevents that the tables which include more variables weight too much in the analysis.

This method can be very useful to analyze surveys for which one can identify several groups of variables, or for which the same questions are asked at several time intervals.

The authors that developed the method (Escofier and Pagès, 1984) particularly insisted on the use of the results which are obtained from the MFA. The originality of method is that it allows visualizing in a two or three dimensional space, the tables (each table being represented by a point), the variables, the principal axes of the analyses of the first phase, and the individuals. In addition, one can study the impact of the other tables on an observation by simultaneously

visualizing the observation described by the all the variables and the projected observations described by the variables of only one table.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Observations/variables table: Select the data that correspond to n observations described by p variables and grouped into K tables. If column headers have been selected, check that the "Variable labels" option has been activated. If the data is **not suitable** for this type of table, it can be **transformed** into a horizontal table, as described above, using the **Create a Products Table.** feature.

Number of tables: Enter the number K of tables in which the selected data are subdivided.

Table labels: Activate this option if you want to use labels for the K tables. If this option is not activated, the name of the tables are automatically generated (Table1, Table2, etc.). If column headers have been selected, check that the "Variable labels" option has been activated.

Number of variables per table:

- **Equal**: Choose this option if the number of variables is identical for all the tables. In that case XLSTAT determines automatically the number of variables in each table.
- **User defined**: Choose this option to select a column that contains the number of variables contained in each table. If the "Variable labels" option has been activated, the first row must correspond to a header.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header. Where the selection is a correlation or covariance matrix, if this option is activated, the first column must also include the variable labels.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated. If your data contains frequency tables, weights will be automatically set to 1 even if you have selected a vector of weights.

Options tab:

Data type: Specify the type of data contained in the various tables, knowing that the type must be the same within a given table. In the case where the "Mixed type" is selected, you need to select a column that indicates the type of data in each table. Use 0 for a table that contains quantitative variables, 1 for a table that contains qualitative variables and 2 for frequency tables.

Filter factors: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %**: Activate this option then enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Activate this option to set the number of factors to take into account.

PCA options: (only for quantitative tables)

- **PCA type**: you can choose between correlation (normalized PCA) or covariance (non normalized PCA).

MCA options: (only for qualitative tables)

- **Sort categories alphabetically**: Activate this option so that the categories of all the variables are sorted alphabetically.

- **Variable-Category labels**: Activate this option to use variable-category labels when displaying outputs. Variable-Category labels include the variable name as a prefix and the category name as a suffix.

Supplementary data tab:

Supplementary observations: Activate this option if you want to calculate the coordinates and represent additional observations. These observations are not taken into account for the factor axis calculations (passive observations as opposed to active observations). If the first row of the data selection for supplementary observations includes a header you must activate the "Variable labels for supp. obs" option. You can also select labels for supplementary observations which will be used for the display.

Supplementary tables: Activate this option if you want to use some tables as supplementary tables. The variables of these tables will not be taken into account for the computation of the factors of the MFA. However, the separate analyses of the first phase of the MFA will be run on these tables. Select a binary vector that let XLSTAT know which are among the K tables the active ones (1) and the supplementary ones (0).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data has been detected.

Remove observations: Activate this option to ignore the observations that contain missing data.

Adapted strategies: Activate this option to choose strategies that are adapted to the data type.

- Quantitative variables:
- **Mean**: Activate this option to estimate the missing data of an observation by the mean of the corresponding variable.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.
- Qualitative variables:
- **New category**: Choose this option to group missing data into a new category of the corresponding variable.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

The outputs tab is divided into five sub-tabs:

General:

These outputs concern all the analyses:

Descriptive statistics: Activate this option to display the descriptive statistics for all the selected variables.

Display results for separate analyses: Activate this option to display results of analysis of each table separately. If the option is not activated only MFA results will be displayed.

Eigenvalues: Activate this option to display the table and chart (scree plot) of eigenvalues.

Contributions: Activate this option to display the contribution tables.

Squared cosines: Activate this option to display the tables of squared cosines.

PCA:

These outputs only concern the PCA:

Factor loadings: Activate this option to display the coordinates of the variables in the factor space.

Variables/Factors correlations: Activate this option to display correlations between factors and variables.

Factor scores: Activate this option to display the coordinates of the observations (factor scores) in the new space created by PCA.

MCA:

These outputs only concern the MCA:

Factor loadings: Activate this option to display the coordinates of the categories in the factor space.

Factor scores: Activate this option to display the coordinates of the observations in the factor space.

CA:

These outputs only concern the CA:

Column coordinates: Activate this option to display the coordinates of the columns in the factor space.

Row coordinates: Activate this option to display the coordinates of the rows in the factor space.

MFA:

These results correspond to the second phase of the MFA:

Tables:

- **Coordinates**: Activate this option to display the coordinates of the tables in the MFA space. Note: the contributions and the squared cosines are also displayed if the corresponding options are checked in the Outputs/General tab.
- **Lg coefficients**: Activate this option to display the Lg coefficients.
- **RV coefficients**: Activate this option to display the RV coefficients.

Variables:

- **Factor loadings**: Activate this option to display the factor loadings in the MFA space.
- **Variables/Factors correlations**: Activate this option to display the correlations between factors and variables in the MFA space.

Partial axes:

- **Maximum number**: Enter the maximum number of factors to keep from the analyses of the first phase that you then want to analyze in the MFA space.
- **Coordinates**: Activate this option to display the coordinates of the partial axes in the space obtained from the MFA.
- **Correlations**: Activate this option to display the correlations between the factors of the MFA and the partial axes.
- **Correlations between axes**: Activate this option to display the correlations between the partial axes.

Observations:

- **Factor scores**: Activate this option to display the factor scores in the MFA space.
- **Coordinates of the projected points**: Activate this option to display the coordinates of the projected points in the MFA space. The projected points correspond to the projections of the observations in spaces reduced to the number of dimensions of each table.

Charts tab:

The charts tab is divided into five sub-tabs:

General:

These options are for all the analyses:

Display charts on two axes: Activate this option if you want the numerous graphical representations displayed after the PCA, MCA and MFA are only displayed on the first two axes, without your being prompted after each analysis.

Options for variables:

Filter: Activate this option to modulate the number of variables displayed:

- **Random**: The observations to display are randomly selected. The "Number of variables" N to display must then be specified.
- **N first variables**: The first N variables are displayed on the chart. The "Number of variables" N to display must then be specified.
- **N last variables**: The last N variables are displayed on the chart. The "Number of variables" N to display must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the variables to display. In the case of display results for separate analyses, this option does not work.
- **Sum(Cos2)>**: Only the variables for which the sum of squared cosines (communalities) are bigger than a value to enter are displayed on the plots.

Options for observations:

Filter: Activate this option to modulate the number of observations displayed:

- **Random**: The observations to display are randomly selected. The "Number of observations" N to display must then be specified.
- **N first rows**: The first N observations are displayed on the chart. The "Number of observations" N to display must then be specified.
- **N last rows**: The last N observations are displayed on the chart. The "Number of observations" N to display must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to display.
- **Sum(Cos2)>**: Only the observations for which the sum of squared cosines (communalities) are bigger than a value to enter will be displayed on the plots.

Color by group: Activate this option, if you want to color observation points according to levels of a qualitative variable. Then select a vertical vector that must have as many rows as there are active observations. If headers were selected for the main table, ensure that a label is also present for the variable in this selection.

- **Confidence ellipses**: Activate this option if you want to display confidence ellipses around group of observations corresponding to the levels of the group variable selected to

color observations. You also have to select the confidence interval for the ellipses.

PCA:

These options concern only the charts of PCA:

Correlation charts: Activate this option to display the charts involving correlations between the components and the variables.

- **Vectors**: Activate this option to display the variables with vectors.

Observations charts: Activate this option to display the charts that allow visualizing the observations in the new space.

- **Labels**: Activate this option to display the observation labels on the charts.

Biplots: Activate this option to display the charts where the input variables and the observations are simultaneously displayed.

- **Vectors**: Activate this option to display the input variables with vectors.
- **Labels**: Activate this option to display the observation labels on the biplots.

Type of biplot: Choose the type of biplot you want to display. See the description section of the PCA for more details.

- **Correlation biplot**: Activate this option to display correlation biplots.
- **Distance biplot**: Activate this option to display distance biplots.
- **Symmetric biplot**: Activate this option to display symmetric biplots.
- **Coefficient**: Choose the coefficient whose square root is to be multiplied by the coordinates of the variables. This coefficient lets you adjust the position of the variable points in the biplot in order to make it more readable. If set to other than 1, the length of the variable vectors can no longer be interpreted as standard deviation (correlation biplot) or contribution (distance biplot).

MCA:

These options concern only the charts of MCA:

Factorial map of categories: Activate this option to display the chart showing the principal coordinates of categories of active and supplementary qualitative variables.

- **Labels**: Activate this option to display the category labels on the charts.

Factorial map of observations: Activate this option to display charts representing the principal coordinates of observations.

- **Labels**: Activate this option to have observation labels displayed on the charts.

Biplot:

Symmetric plots: Activate this option to display on the same chart principal coordinates of observations and principal coordinates of categories.

CA:

These options concern only the charts of CA:

Columns chart: Activate this option to display the chart showing the principal coordinates of columns.

Rows chart: Activate this option to display the chart showing the principal coordinates of rows.

- **Labels**: Activate this option to have observation labels displayed on the charts.

Biplots: Activate this option to display principal coordinates of rows and principal coordinates of columns on the same chart.

MFA:

These options concern only the results of the second phase of the MCA:

Table charts: Activate this option to display the charts that allow to visualize the tables in the MFA space.

Correlation charts: Activate this option to display the charts involving correlations between the components and the quantitative variables used in the MFA.

Observations charts: Activate this option to display the chart of the observations in the MFA space.

Correlation charts (partial axes): Activate this option to display the correlation chart for the partial axes obtained from the first phase of the MFA.

Charts of the projected points: Activate this option to display the chart that shows at the same time the observations in the MFA space, and the observations projected in the sub-space of each table.

- **Observation labels**: Activate this option to display the observations labels on the charts.
- **Projected points labels**: Activate this option to display the labels of the projected points.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. This includes the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

The results of the analyses performed on each individual table (PCA, MCA or CA) are then displayed if you have selected the option *Display results for separate analyses*. These results

are identical to those you would obtain after running the PCA, MCA or CA function of XLSTAT.

Afterwards, the results of the second phase of the MFA are displayed.

First, eigenvalues, variable results and observation results are displayed as in the case of PCA.

At the end of the observations coordinates table, the following button is displayed:  . Click on this button to automatically open the pre-filled dialog box of HAC (Hierarchical Ascending Classification) and perform a classification of the observations on the factorial coordinates.

Then the results specific to MFA are displayed:

The **coordinates of the tables** are then displayed and used to create the plots of the tables. The latter allow to visualize the distance between the tables. The coordinates of the supplementary tables are displayed in the second part of the table. Then contributions and squared cosines for tables are displayed.

Lgcoefficients: The Lg coefficients of relationship between the tables allow to measure to what extent the tables are related two by two. The more variables of a first table are related to the variables of the second table, the higher the Lg coefficient.

RV coefficients: The RV coefficients of relationship between the tables are another measure derived from the Lg coefficients. The value of the RV coefficients varies between 0 and 1.

The **coordinates of the partial axes**, and even more their correlations, allow to visualize in the new space the link between the factors obtained from the first phase of the MFA, and those obtained from the second phase.

The **correlations between partial axes** allow to understand the link between factorial axes of the different analyses.

Last, the **coordinates of the projected points** in the space resulting from the MFA are displayed. The projected points correspond to projections of the observations in the spaces reduced to the dimensions of each table. The representation of the projected points superimposed with those of the complete observations makes it possible to visualize at the same time the diversity of the information brought by the various tables for a given observation, and to visualize the relative distances from two observations according to the various tables.

*Remark about the **Axes homogeneity index***: This index developed by our team is very useful to determine if the contributions of the observations are homogeneous for the different axes. It is constructed as the proportion of observations with an absolute contribution > 1/n. An index above 0.4 indicates a very good homogeneity with well represented observations. On the other hand, an index lower than 0.1 should be a warning to the user who should check if there are no outliers in the variables constructing the axis that would distort its interpretation (the outliers would then be the observations that stand out from the others on the axis in question).

Example

An example of Multiple Factor Analysis on mixed data is available on the XLSTAT Help Center:

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

An example of Multiple Factor Analysis on frequency tables is available on the XLSTAT Help Center:

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

References

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

Use STATIS to analyze multiple configurations of objects / quantitative variables. This method allows you to:

- Study and visualize the links between the objects
- Study the agreements between the configurations

In this section:

Description

The STATIS method is a multi-configurations data analysis method most used in sensometry. The configurations can be various assessors, subjects or judges. STATIS can be particularly used in the case of projective mapping / Napping, conventional profiling, free choice profiling. The great interest of STATIS is that atypical configurations have a smaller weight than this of central configurations. Therefore, the analysis best reflects the general point of view and not those atypical configurations.

Uses of STATIS

There are several applications for STATIS, including:

- Study and visualization of objects in the main plans.
- Study of the links between the configurations, especially to find the most atypicals.

Principle of STATIS

STATIS is a method working on the scalar matrix of each configuration, thus making it possible to work with configurations with different numbers of columns. Its aim is to form a consensus configuration that reflects at best the different configurations. This consensus can then be projected on different axes. If the information associated with 2 or 3 first axes represents a sufficient percentage of the total variability of the consensus, the objects will be able to be represented on a 2- 3-dimensional chart, thus making interpretation much easier.

Structure of the data

Two cases exist:

1. The number of the p variables is identical for the m configurations.

2. The number p of the variables varies from one configuration to the other.

For data entry, XLSTAT asks you to select a configuration corresponding to the m contiguous configurations, and to give the case of structure. If the data is **not suitable** for this type of table, it can be **transformed** into a horizontal table, as described above, using the **Create a Products Table.** feature.

Scaling and global scaling

If the data within a configuration are not on the same scale, it is advisable to scale (reduce) the variables of each configuration. For example, this is not the case for ratings between 0 and 20, but it is advised if some notes lie between 0 and 10 and others between 0 and 20.

Classically, the overall reduction of each configuration is recommended. It allows to put all the configurations on an equal footing in term of variance. For example, in the case, of configurations where the attributes are noted between 0 and 20 by assessors, it will remove scale factors between the assessor that only notes between 5 and 15 and the assessor that uses the full scale of notes.

Interpreting the results

The representation of objects in the space of k factors allows you to visually interpret the proximities between the objects, by means of precautions.

We can consider that the projection of an object on a plan is reliable if the object is far from the center of the graph.

Number of factors

Two methods are commonly used to determine how many factors must be retained for the interpretation of the results:

- Watch the decreasing curve of eigenvalues. The number of factors to be kept corresponds to the first turning point found on the curve.

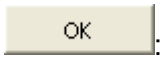
- We can also use the cumulative variability percentage represented by the factor axes and decide to use only a certain percentage.

Graphic representations

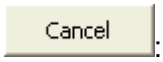
However, these representations are only reliable if the sum of the variability percentages associated with the axes of the representation space are sufficiently high. If this percentage is high (for example 80%), the representation can be considered as reliable. If the percentage is low, it is recommended to produce representations on several axis pairs in order to validate the interpretation made on the two first factor axes.

Dialog box

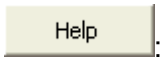
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Configurations: Select the data that correspond to the configurations. If a column header has been selected, check that the "Variable labels" option has been activated. If the data is **not suitable** for this type of table, it can be **transformed** into a horizontal table, as described above, using the **Create a Products Table.** feature.

Number of configurations: Enter the number of contiguous configurations in the configurations table.

Number of variables per configuration:

- **Equal**: Choose this option if the number of variables is identical for all the configurations. In that case XLSTAT determines automatically the number of variables in each configuration.
- **User defined**: Choose this option to select a column that contains the number of variables contained in each configuration. If the "Variable labels" option has been

activated, the first row must correspond to a header.

Configuration labels: Check this option if you want to use the available configuration labels. If you don't check this option, labels will be created automatically (Config.1, Config.2, etc.). If a column header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row (or column if in transposed mode) of the selected data (configurations, configuration labels, object labels) contains a header.

Object labels: Check this option if you want to use the available configuration labels. If you do not check this option, labels will be created automatically (Obj.1, Obj. 2, etc.). If a column header has been selected, check that the "Variable labels" option has been activated.

Options tab:

Scaling (variables): Activate this option to reduce the variables.

Global scaling: Activate this option to globally reduce the configurations.

Filter factors: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %**: Enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Set the number of factors to take into account.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Estimate missing data: Activate this option to estimate the missing data before the calculation starts.

- **Mean or mode**: Activate this option to estimate the missing data by using the mean (quantitative variables) or the mode (qualitative variables) for the corresponding variables.

- **Nearest neighbor**: Activate this option to estimate the missing data for an observation by searching for the nearest neighbor to the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Eigenvalues: Activate this option to display the table and chart (scree plot) of eigenvalues.

Consensus coordinates: Activate this option to display the coordinates of the consensus in the factors space.

RV matrix: Activate this option to display the RV matrix.

Scaling factors: Activate this option to display the scaling factors.

Weights: Activate this option to display the weights created and used by STATIS.

Consensus configuration: Activate this option to display the consensus configuration created by STATIS.

Homogeneity: Activate this option to display homogeneity of the configurations.

RV config/consensus: Activate this option to display the RV coefficient between each configuration and the consensus.

Global error: Activate this option to display the error of the STATIS criterion.

Residuals by configuration: Activate this option to display the error of the STATIS criterion for each configuration.

Residuals by object: Activate this option to display the error of the STATIS criterion for each object.

Correlations: Activate this option to display the correlations between factors and the initial variables.

Coordinates of the projected points: Activate this option to display the coordinates of the projected points in the factor space. The projected points correspond to the projections of the objects of each configuration in the factor space.

- **Presentation by configuration**: Activate this option to display a table coordinates by configuration.
- **Presentation by object**: Activate this option to display a table coordinates by object.

Charts tab:

Display charts on two axes: Activate this option if you want the different graphical representations to be displayed only on the first two axes.

Eigenvalues: Activate this option to display the *scree plot* of the eigenvalues.

Consensus coordinates: Activate this option to display the plot of the consensus coordinates in the factors space.

Scaling factors: Activate this option to display the bar chart of the scaling factors.

Weights: Activate this option to display the bar chart of the weights created and used by STATIS.

RV config/consensus: Activate this option to display the bar chart of the RV coefficient between each configuration and the consensus.

Residuals by configuration: Activate this option to display the bar chart of the error of the STATIS criterion for each configuration.

Residuals by object: Activate this option to display the bar chart of the error of the STATIS criterion for each object.

Correlations: Activate this option to display charts showing the correlations between the factors and initial variables. This chart is named correlation circle.

Charts of the projected points: Activate this option to display the graphic representing both the objects, and the objects of each of the configurations projected in the factor space.

- **Observation labels**: Activate this option to display the observations labels on the charts.
- **Projected points labels**: Activate this option to display the labels of the projected points.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. This includes the number of observations, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

Eigenvalues: The eigenvalues and corresponding chart (*scree plot*) are displayed.

Consensus coordinates: Consensus coordinates in the factors space are displayed, with the corresponding charts (depending on the number of factors chosen).

RV matrix: The matrix of RV coefficients between all configurations is displayed. The coefficient RV is a coefficient of similarity between two configurations included between 0 and 1. The closer it is to 1, the stronger the similarity. This matrix is used by STATIS to calculate the weights of the configurations.

Scaling factors: The scaling factors are displayed with the associated bar chart. The larger a scale factor in a configuration, the smaller the scale of the configuration used. This table is used in sensory analysis to understand how assessors use differently the rating scales.

Weights: The weights calculated by STATIS are displayed, with the associated bar chart. The greater the weight, the more the configuration contributed to the consensus. Knowing that STATIS gives more weight to the closest configurations from a global point of view, a much lower weight than the others will mean that the configuration is atypical.

Consensus configuration: The consensus configuration is displayed. It corresponds to the weighted average of the scalar product matrices of the initial configurations (possibly reduced by variable and / or globally).

Homogeneity: The homogeneity of the configurations is displayed. It is a value between 1/m (m being the number of configurations) and 1, which increases with the homogeneity of the configurations.

RV config/consensus: The RV coefficients between the configurations and the consensus are displayed, with the associated bar chart. Like the weights of STATIS, these coefficients make it possible to detect atypical configurations. The advantage of these coefficients is that they are between 0 and 1, so they are easier to interpret than the weights.

Global error: The global error of the STATIS criterion is displayed. It corresponds to the sum of all residuals (which can be presented by configuration or object).

Residuals by configuration: This table and the corresponding bar chart make it possible to visualize the distribution of the residuals by configuration. It is thus possible to identify for which configurations STATIS has been less efficient, or in other words, which configurations stand out the most from the consensus.

Residuals by object: This table and the corresponding bar chart make it possible to visualize the distribution of the residuals by object. It is thus possible to identify for which objects STATIS has been less efficient, or in other words, which objects stand out the most from the consensus.

Correlations: The correlations between the factors and the initial variables, as well as the correlation circle are displayed. This chart shows the links between the different variables and the factors.

Objects coordinates (presentation by configuration): This series of tables corresponds to the coordinates of the objects for each configuration, after the optional scaling and global scaling then the projection on the factors. The presentation is made by configuration.

Objects coordinates (presentation by object): This series of tables corresponds to the coordinates of the objects for each configuration, after the optional scaling and global scaling then the projection on the factors. The presentation is made by object.

Coordinates of the projected points: The projected points correspond to the projections of the objects of each configuration in the factor space. The representation of the projected points superimposed with those of the objects makes it possible to visualize at the same time the diversity of the information brought by the various configurations for a given object, and to visualize the relative distances from two objects according to the various configurations.

Example

A tutorial on how to use STATIS is available on the XLSTAT Help Center:

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

References

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

Use CLUSTATIS to build homogeneous classes of configurations/tables. In the context of sensory analysis, this function allows you to perform a cluster analysis of subjects on the basis of their perceptions of products.

In this section:

Description

Cases where data are made up of different blocks of variables are becoming more and more frequent. Sensory analysis is particularly concerned by this phenomenon, since many tasks lead to this type of data, with each consumer/judge/subject providing a configuration/table (*e.g.* Projective mapping/Napping, conventional profiling, free choice profiling). As perceptions between subjects are often different, a clustering of the subjects may be necessary. The CLUSTATIS method fits into this context. Moreover, this strategy allows to set aside configurations that do not conform to any of the constructed classes, which correspond to atypical subjects in the framework of sensory analysis.

Principle of CLUSTATIS

CLUSTATIS is a clustering method based on the matrices of the scalar products of each configuration, which allows to consider configurations with different numbers of columns. The objective of this method is to constitute classes of configurations that are as homogeneous as possible, each group of configurations being represented by a latent configuration (called consensus) determined by STATIS. It is therefore natural that each class is finally analysed by STATIS, in order to determine the differences between the constituted classes. CLUSTATIS consists of a hierarchical algorithm that can be "consolidated" by a partitioning algorithm (*i.e.* the partitioning algorithm is initialized by cutting the dendrogram). An interesting option is the creation of a class "K+1" (corresponding to an additional class) in order to set aside tables that do not conform to any class. A configuration will be placed in this class if the similarities (RV coefficients) between the consensus of each class and this configuration are all considered weak.

Structure of the data

Two cases exist:

1. The number of the p variables is identical for the m configurations.

2. The number p of the variables varies from one configuration to the other.

For data entry, XLSTAT asks you to select a configuration corresponding to the m contiguous configurations, and to give the case of structure. If the data is **not suitable** for this type of table, it can be **transformed** into a horizontal table, as described above, using the **Create a Products Table.** feature.

Scaling

If the data within a configuration are not on the same scale, it is advisable to scale (reduce) the variables of each configuration. For example, this is not the case for ratings between 0 and 20, but it is advised if some notes lay between 0 and 10 and others between 0 and 20.

Interpreting the results

For each class, the representation of objects in the space of factors allows to visually interpret the proximities between the objects, by means of precautions. We can consider that the projection of an object on a plan is reliable if the object is far from the center of the graph.

Since the class "K+1" contains tables that do not conform to any of the classes, this class is very dependent on the number of groups.

Number of factors

Two methods are commonly used to determine how many factors must be retained for the interpretation of the results:

- Watch the decreasing curve of eigenvalues. The number of factors to be kept corresponds to the first turning point found on the curve.

- We can also use the cumulative variability percentage represented by the factor axes and decide to use only a certain percentage.

Graphic representations

The graphical representations of the objects in each class are only reliable if the sum of the variability percentages associated with the axes of the representation space are sufficiently high. If this percentage is high (for example 80%), the representation can be considered as reliable. If the percentage is low, it is recommended to produce representations on several axis pairs in order to validate the interpretation made on the two first factor axes.

Quality of a cluster analysis

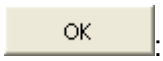
In order to determine the quality of a hierarchical clustering, one can use the increase in within-class variance (CLUSTATIS criterion error) caused by the merging of two classes. This increase is equal to the height of the dendrogram in which the two classes of configurations are grouped in the same class.

The homogeneity of each class and the global homogeneity are also very important indices (between $1/m$ and 1, m being the number of configurations) which allow to judge the quality of

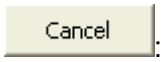
the cluster analysis. It should be noted that the consolidation and the addition of a class "K+1" can increase homogeneities.

Dialog box

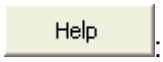
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.


: Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Configurations: Select the data that correspond to the configurations. If a column header has been selected, check that the "Variable labels" option has been activated. If the data is **not suitable** for this type of table, it can be **transformed** into a horizontal table, as described above, using the **Create a Products Table.** feature.

Number of configurations: Enter the number of contiguous configurations in the configurations table.

Number of variables per configuration:

- **Equal**: Choose this option if the number of variables is identical for all the configurations. In that case XLSTAT determines automatically the number of variables in each configuration.

- **User defined**: Choose this option to select a column that contains the number of variables contained in each configuration. If the "Variable labels" option has been activated, the first row must correspond to a header.

Configuration labels: Check this option if you want to use the available configuration labels. If you don't check this option, labels will be created automatically (Config.1, Config.2, etc.). If a column header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row (or column if in transposed mode) of the selected data (configurations, configuration labels, object labels) contains a header.

Object labels: Check this option if you want to use the available configuration labels. If you do not check this option, labels will be created automatically (Obj.1, Obj. 2, etc.). If a column header has been selected, check that the "Variable labels" option has been activated.

Options tab:

Scaling (variables): Activate this option to reduce the variables.

Filter factors: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %**: Enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Set the number of factors to take into account.

Truncation: Activate this option is you want XLSTAT to **automatically** define the truncation level, and therefore the number of classes to retain, or if you want to define the **number of classes** to create, or the **level** at which the dendrogram is to be truncated.

Consolidation: Activate this option to perform a consolidation of the classes obtained from the dendrogram.

Class K+1: Activate this option to add an additional class that will contain the configurations that do not fit any pattern of classes.

Rho parameter: Choose how you want to set the rho parameter: **automatically** or **user-defined**. This parameter represents the minimum agreement to be considered as sufficiently in agreement to be kept in a class. The higher this parameter is set, the stronger the agreement required with the class and the more likely you are to place configurations in the K+1 class.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Estimate missing data: Activate this option to estimate the missing data before the calculation starts.

- **Mean or mode**: Activate this option to estimate the missing data by using the mean (quantitative variables) or the mode (qualitative variables) for the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data for an observation by searching for the nearest neighbor to the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

RV matrix: Activate this option to display the matrix of RV coefficients between configurations.

Node statistics: Activate this option to display the statistics for dendrogram nodes.

Class composition: Activate this option to display the composition of each class.

Eigenvalues: Activate this option to display for each class the table and chart (scree plot) of eigenvalues.

Consensus coordinates: Activate this option to display for each class the coordinates of the consensus in the factor space.

Consensus configuration: Activate this option to display the consensus configuration of each class created by STATIS.

RV config/consensus: Activate this option to display the RV coefficient between each configuration and the consensus of its class.

Weights: Activate this option to display the weights created and used by STATIS in each class.

Homogeneities: Activate this option to display the homogeneity of each class as well as the global homogeneity.

Global Error/Within-Class variance: Activate this option to display the error of the CLUSTATIS minimization criterion, equivalent to the within-class variance.

RV between consensus: Activate this option to display the RV coefficient between each consensus configuration.

Charts tab:

Levels bar chart: Activate this option to display the diagram of levels showing the impact of successive clusterings on the within-class variance.

Dendrogram: Activate this option to display the dendrogram.

- **Horizontal**: Choose this option to display a horizontal dendrogram.
- **Vertical**: Choose this option to display a vertical dendrogram.
- **Truncated**: Activate this option to display the truncated dendrogram (the dendrogram starts at the level of the truncation).
- **Labels**: Activate this option to display configuration labels (full dendrogram) or classes (truncated dendrogram) on the dendrogram.
- **Colors**: Activate this option to use colors to represent the different groups on the full dendrogram.

Display charts on two axes: Activate this option if you want the different graphical representations to be displayed only on the first two axes.

Eigenvalues: Activate this option to display the *scree plot* of the eigenvalues.

Consensus coordinates: Activate this option to display the plot of the consensus coordinates in the factors space for each class.

RV config/consensus: Activate this option to display the bar chart of the RV coefficient between each configuration and the consensus of its class.

Weights: Activate this option to display the bar chart of the weights created and used by STATIS.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. This includes the number of observations (objects), the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

RV matrix: The matrix of RV coefficients between all configurations is displayed. The RV coefficient is an index of similarity between two configurations included between 0 and 1. The closer it is to 1, the stronger the similarity.

Node statistics: This table shows the data for the successive nodes in the dendrogram. The first node has an index which is the number of configurations increased by 1. Thus, it is easy to see at any time if a configuration or group of configurations is clustered with another group of configurations in the dendrogram.

Levels bar chart: This table displays the statistics for dendrogram nodes, which correspond to the increase in the CLUSTATIS minimization criterion (equivalent to the increase in within-class variance) when merging two classes.

Dendrograms: The full dendrogram displays the progressive clustering of configurations. If truncation has been requested, a broken line marks the level the truncation has been carried out. The truncated dendrogram shows the classes after truncation.

Composition of classes:

Results by configuration: This table shows the assignment class for each configuration in the initial configuration order. If a consolidation is requested, the results are given before and after the consolidation. If you have checked "class K+1", it is possible that some tables may have a missing value after consolidation. This means that they are not placed in any of the main classes (they are placed in class "K+1").

Results by class: The results are given by class. Thus, a list of configurations is displayed for each class.

Number of configurations per class: The number of configurations in each class is indicated.

Rho parameter computed: Result displayed only if you have chosen to add a class "K+1". The rho parameter represents the minimum similarity that a configuration must have with the consensus of a class in order to belong to it. If this condition is not met for any of the classes, the configuration is placed in class "K+1". This parameter is calculated according to the proximity of each configuration to its class as well as to the neighboring class.

Analysis of the class k:

In this section, the analysis of each of the classes by the STATIS method is displayed.

Eigenvalues: The eigenvalues and corresponding chart (*scree plot*) are displayed.

Consensus coordinates: Consensus coordinates in the factors space are displayed, with the corresponding charts (depending on the number of factors chosen).

Consensus configuration: The consensus configuration is displayed. It corresponds to the weighted average of the scalar product matrices of the initial configurations (reduced globally and possibly reduced by variable).

RV config/consensus: The RV coefficients between the configurations and the consensus are displayed, with the associated bar chart. Like the weights of STATIS, these coefficients allow to detect atypical configurations. The advantage of these coefficients is that they are between 0 and 1, so they are easier to interpret than the weights.

Weights: The weights calculated by STATIS are displayed, with the associated bar chart. The greater the weight, the more the configuration contributed to the consensus. Knowing that STATIS gives more weight to the closest configurations from a global point of view, a much lower weight than the others will mean that the configuration is atypical.

Indices:

Homogeneities: The homogeneity of each class is displayed. It is a value between $1/m$ (m being the number of configurations of the class) and 1, which increases with the homogeneity of the configurations. In a second step, the global homogeneity, which is a weighted average of the homogeneity of each class, is displayed.

Global Error/Within-class Variance: The error of the CLUSTATIS criterion is displayed. It corresponds to the within-class variance.

RV between consensus : The matrix of the RV coefficients between the consensus of each class is displayed. This matrix shows how close the classes are to each other.

Example

A tutorial on how to use CLUSTATIS is available on the XLSTAT Help Center:

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

References

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

Use CATATIS to analyze Check-All-That-Apply (CATA) data. This method allows:

- To study and visualize the links between the products and attributes.
- To study the agreements between the assessors.

In this section:

Description

The CATATIS method is an improvement of the usual method for processing CATA data. It is considered as the equivalent of the STATIS method for this type of data. The great interest of CATATIS lies in the fact that atypical assessors have a lower weight than those who agree with the rest of the panel. Therefore, the analysis best reflects the general point of view and not those atypical assessors.

In addition, panel consistency tests overall and by attribute are proposed to determine if certain attributes are not included. Tests on the weights to determine if some subjects have a non-significant weight can also be computed. This last option is particularly useful when dealing with experts.

The fact that the subjects have had several sessions is allowed. In this case, then the average is used per subject (if for a given product and a given attribute they have checked once and unchecked the other time, the average of 0.5 will be taken into account). Moreover, non-binary data are accepted and the repeatability from one session to another is controlled.

Uses of CATATIS

There are several applications for CATATIS, including:

- Study and visualization of the products and the attributes on the main planes.
- Study of the links between the assessors, especially to find the most atypical ones.

Principle of CATATIS

The goal of CATATIS is to form a consensus configuration that reflects at best the different assessors. This consensus can then be projected on different axes by a Correspondence Analysis. If the information associated with 2 or 3 first axes represents a sufficient percentage of the total variability of the consensus, the products and attributes will be able to be represented on a 2- 3-dimensional chart, thus making interpretation much easier.

Structure of the data

There are two different formats:

1. All the data are merged horizontally (horizontal format).
2. All the data are merged vertically (vertical format).

For data entry, XLSTAT asks you to select all the data, and to give the format type. In the case of the vertical format, product and assessor labels are mandatory.

Interpreting the results

The representation of the products and attributes in the space of k factors allows to visually interpret the proximities between the products and attributes, by means of precautions.

We can consider that the projection of a product or an attribute on a plan is reliable if it is far from the center of the graph.

Number of factors

Two methods are commonly used to determine how many factors must be retained for the interpretation of the results:

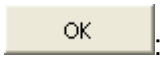
- Watch the decreasing curve of eigenvalues. The number of factors to be kept corresponds to the first turning point found on the curve.
- We can also use the cumulative variability percentage represented by the factor axes and decide to use only a certain percentage.

Graphic representations

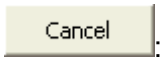
These representations are only reliable if the sum of the variability percentages associated with the axes of the representation space are sufficiently high. If this percentage is high (for example 80%), the representation can be considered as reliable. If the percentage is low, it is recommended to produce representations on several axis pairs in order to validate the interpretation made on the two first factor axes.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.


: Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

CATA data (0/1): Select the data that correspond to the different assessors. If a column header has been selected, check that the "Attribute labels" option has been activated.

Format: Click on horizontal or vertical depending on the structure of your data.

If the format is **horizontal**:

Number of assessors: Enter the number of assessors in CATA data.

Product labels: Check this option if you want to use the available product labels. If you do not check this option, labels will be created automatically. If a column header has been selected, check that the "Attribute labels" option has been activated.

Assessor labels: Check this option if you want to use the available assessor labels. If you do not check this option, labels will be created automatically. If a column header has been selected, check that the "Attribute labels" option has been activated.

If the format is **vertical**:

Products: Select the products corresponding to the CATA data rows. If a column header has been selected, check that the "Attribute labels" option has been activated.

Assessors: Select the assessors corresponding to the CATA data rows. If a column header has been selected, check that the "Attribute labels" option has been activated.

Sessions: select the sessions corresponding to the rows of the CATA data. If a column header has been selected, check that the "Attribute labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Attribute labels: Activate this option if the first row (or column if in transposed mode) of the selected data (CATA data (0/1), Product labels, Assessor labels) contains a header.

Options tab:

Filter factors: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %**: Enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Set the number of factors to take into account.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Replace missing data by 0: Activate this option to replace missing data by 0.

Outputs tab:

Descriptive statistics: Activate this option to display the number of checks by each assessor.

CA eigenvalues: Activate this option to display the table and chart (scree plot) of eigenvalues of the CA on the consensus.

CA coordinates: Activate this option to display the coordinates of the consensus in the factors space.

Similarity matrix (S): Activate this option to display the similarity matrix (Ochiai index).

Scaling factors: Activate this option to display the scaling factors.

Weights: Activate this option to display the weights created and used by CATATIS.

Weight tests: Activate this option to test if the weights of the subjects are significant.

Consensus configuration: Activate this option to display the consensus configuration created by CATATIS.

Homogeneity: Activate this option to display homogeneity of the assessors.

Consistency tests: Activate this option to test if the panel is consistent globally and by attribute.

Similarity assessors/consensus: Activate this option to display the similarity coefficient between each assessor and the consensus.

Global error: Activate this option to display the error of the CATATIS criterion.

Residual per assessor: Activate this option to display the error of the CATATIS criterion for each assessor.

Residual per product: Activate this option to display the error of the CATATIS criterion for each product.

Charts tab:

CA eigenvalues: Activate this option to display the *scree plot* of the CA eigenvalues.

CA biplot: Activate this option to display the plot of the consensus coordinates in the factors space.

Display charts on two axes: Activate this option so that XLSTAT does not prompt you to select the axes, and automatically displays the biplot on the first two axes.

Scaling factors: Activate this option to display the bar chart of the scaling factors.

Weights: Activate this option to display the bar chart of the weights created and used by CATATIS.

Similarity assessors/consensus: Activate this option to display the bar chart of the similarity index between each assessor and the consensus.

Residual per assessor: Activate this option to display the bar chart of the error of the CATATIS criterion for each assessor.

Residual per product: Activate this option to display the bar chart of the error of the CATATIS criterion for each product.

Results

Assessors' repeatability: The coefficient of similarity (Salton Cosine) between the results of different sessions is displayed. This coefficient takes values between 0 and 1 and increases

with the similarity between sessions.

Descriptive statistics: The number of checks by assessor is displayed. Warning, if you have entered non-binary data, this number may be decimal.

Eigenvalues of CA: The eigenvalues of CA and corresponding chart (*scree plot*) are displayed.

Product coordinates: The coordinates of the products of the consensus in the factors space are displayed, with the corresponding charts (depending on the number of chosen factors).

Attribute coordinates: The coordinates of the attributes of the consensus in the factors space are displayed, with the corresponding charts (depending on the number of chosen factors).

Similarity matrix (S): The matrix of similarity index between all assessors is displayed. The similarity index is included between 0 and 1. The closer it is to 1, the stronger the similarity. This matrix is used by CATATIS to calculate the weights of the assessors.

Scaling factors: The scaling factors are displayed with the associated bar chart. The larger a scale factor of an assessor, the smaller the number of checks of this assessor.

Weights: The weights calculated by CATATIS are displayed, with the associated bar chart. The greater the weight, the more the assessor contributed to the consensus. Knowing that CATATIS gives more weight to the closest assessor from a global point of view, a much lower weight than the others will mean that the assessor is atypical.

Weight tests: The results of the weight tests are displayed. If a subject has a non-significant weight, then his point of view is very different from the global point of view, and his results can be questioned if he is an expert.

Consensus configuration: The consensus configuration is displayed. It corresponds to the weighted average of the initial data.

Homogeneity: The homogeneity of the assessors is displayed. It is a value between $1/m$ (m being the number of assessors) and 1, which increases with the homogeneity of the assessors.

Consistency tests: The results of the consistency tests are displayed globally and by attribute. If the panel is globally inconsistent, the data can unfortunately be discarded. If it is inconsistent for one or more attributes, then those attributes are subject to so much disagreement that they have surely been misunderstood.

Distance between the median of permutations and homogeneity: This distance shows how strong the homogeneity of subjects is compared to random responses.

Similarity assessors/consensus: The similarity indices between the assessors and the consensus are displayed, with the associated bar chart. Like the weights of CATATIS, these coefficients make it possible to detect atypical assessors. The advantage of these coefficients is that they are between 0 and 1, so they are easier to interpret than the weights.

Global error: The global error of the CATATIS criterion is displayed. It corresponds to the sum of all residuals (which can be presented by assessor or product).

Residual per assessor: This table and the corresponding bar chart make it possible to visualize the distribution of the residuals by assessor. It is thus possible to identify for which

assessors CATATIS has been less efficient, or in other words, which assessors stand out the most from the consensus.

Residual per product: This table and the corresponding bar chart make it possible to visualize the distribution of the residuals by product. It is thus possible to identify for which products CATATIS has been less efficient, or in other words, which products stand out the most from the consensus.

Example

A tutorial on how to use CATATIS is available on the XLSTAT Help Center:

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

References

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. (2020). Classification de tableaux de données, applications en analyse sensorielle (Doctoral dissertation, Nantes, Ecole nationale vétérinaire).

Llobell, F., Bonnet, L., & Giacalone, D. (2024). Assessment of panel performance in CATA and RATA experiment. Journal of Sensory Studies, 39(4), e12941.

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.

CLUSCATA

Use CLUSCATA to build homogeneous classes of judges based on their perceptions of products.

In this section:

Description

CATA tasks are widely used nowadays. However, product perceptions often differ among assessors. A cluster analysis of the assessors may therefore be necessary. The CLUSCATA method fits into this context. Moreover, this strategy allows to put aside the assessors who do not fir any pattern of classes. CLUSCATA can be seen as an adaptation of CLUSTATIS for CATA data.

Principle of CLUSCATA

The objective of CLUSCATA is to constitute classes of assessors as homogeneous as possible, each class of assessors being represented by a latent table (called consensus) determined by CATATIS. It is therefore natural that each class is finally analyzed by CATATIS, in order to determine the differences between the constituted classes. CLUSCATA consists of a hierarchical algorithm that can be "consolidated" by a partitioning algorithm (*i.e.* the partitioning algorithm is initialized by cutting the dendrogram). An interesting option is the creation of a "K+1" class (corresponding to an additional class) in order to set aside assessors who do not conform to any class. An assessor will be placed in this class if the similarities (Ochiai coefficients) between the consensus of each class and this assessor are all considered weak.

Structure of the data

There are two different formats:

1. All the data are merged horizontally (horizontal format).
2. All the data are merged vertically (vertical format).

For data entry, XLSTAT asks you to select all the data, and to give the format type. In the case of the vertical format, product and assessor labels are mandatory.

Interpreting the results

The representation of the products and attributes in the space of k factors allows to visually interpret the proximities between the products and attributes, by means of precautions.

We can consider that the projection of a product or an attribute on a plan is reliable if it is far from the center of the graph.

Number of factors

Two methods are commonly used to determine how many factors must be retained for the interpretation of the results:

- Watch the decreasing curve of eigenvalues. The number of factors to be kept corresponds to the first turning point found on the curve.

- We can also use the cumulative variability percentage represented by the factor axes and decide to use only a certain percentage.

Graphic representations

The graphical representations of the objects in each class are only reliable if the sum of the variability percentages associated with the axes of the representation space are sufficiently high. If this percentage is high (for example 80%), the representation can be considered as reliable. If the percentage is low, it is recommended to produce representations on several axis pairs in order to validate the interpretation made on the two first factor axes.

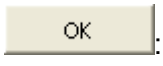
Quality of a cluster analysis

In order to determine the quality of a hierarchical clustering, one can use the increase in within-class variance (CLUSCATA criterion error) caused by the merging of two classes. This increase is equal to the height of the dendrogram in which the two classes of assessors are grouped in the same class.

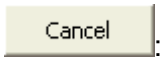
The homogeneity of each class and the global homogeneity are also very important indices (between $1/m$ and 1, m being the number of assessors) which allow to judge the quality of the cluster analysis. It should be noted that the consolidation and the addition of a class "K+1" can increase homogeneities.

Dialog box

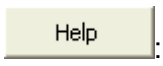
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

CATA data (0/1): Select the data that correspond to the different assessors. If a column header has been selected, check that the "Attribute labels" option has been activated.

Format: Click on horizontal or vertical depending on the structure of your data.

If the format is **horizontal**:

Number of assessors: Enter the number of assessors in CATA data.

Product labels: Check this option if you want to use the available product labels. If you do not check this option, labels will be created automatically. If a column header has been selected, check that the "Attribute labels" option has been activated.

Assessor labels: Check this option if you want to use the available assessor labels. If you do not check this option, labels will be created automatically. If a column header has been selected, check that the "Attribute labels" option has been activated.

If the format is **vertical**:

Products: Select the products corresponding to the CATA data rows. If a column header has been selected, check that the "Attribute labels" option has been activated.

Assessors: Select the assessors corresponding to the CATA data rows. If a column header has been selected, check that the "Attribute labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Attribute labels: Activate this option if the first row (or column if in transposed mode) of the selected data (CATA data (0/1), Product labels, Assessor labels) contains a header.

Options tab:

Filter factors: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %**: Enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Set the number of factors to take into account.

Truncation: Activate this option is you want XLSTAT to **automatically** define the truncation level, and therefore the number of classes to retain, or if you want to define the **number of classes** to create, or the **level** at which the dendrogram is to be truncated.

Consolidation: Activate this option to perform a consolidation of the classes obtained from the dendrogram.

Class K+1: Activate this option to add an additional class that will contain the assessors that do not fit any pattern of classes.

Rho parameter: Choose how you want to set the rho parameter: **automatically** or **user-defined**. This parameter represents the minimum agreement to be considered as sufficiently in agreement to be kept in a class. The higher this parameter is set, the stronger the agreement required with the class and the more likely you are to place assessors in the K+1 class.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Replace missing data by 0: Activate this option to replace missing data by 0.

Outputs tab:

Descriptive statistics: Activate this option to display the number of checks by each assessor.

Similarity matrix (S): Activate this option to display the similarity matrix (Ochiai index).

Node statistics: Activate this option to display the statistics for dendrogram nodes.

Class composition: Activate this option to display the composition of each class.

CA eigenvalues: Activate this option to display the table and chart (scree plot) of eigenvalues of the CA on the consensus of each class.

CA coordinates: Activate this option to display the coordinates of the consensus of each class in the factors space.

Consensus configuration: Activate this option to display the consensus configuration of each class created by CATATIS.

Similarity assessors/consensus: Activate this option to display the similarity coefficient between each assessor and the consensus of its class.

Weights: Activate this option to display the weights of the assessors created and used by CATATIS in each class.

Homogeneities: Activate this option to display the homogeneity of each class as well as the global homogeneity.

Global Error/Within-Class variance: Activate this option to display the error of the CLUSCATA minimization criterion, equivalent to the within-class variance.

Charts tab:

Levels bar chart: Activate this option to display the diagram of levels showing the impact of successive clusterings on the within-class variance.

Dendrogram: Activate this option to display the dendrogram.

- **Horizontal**: Choose this option to display a horizontal dendrogram.
- **Vertical**: Choose this option to display a vertical dendrogram.
- **Truncated**: Activate this option to display the truncated dendrogram (the dendrogram starts at the level of the truncation).
- **Labels**: Activate this option to display assessors labels (full dendrogram) or classes (truncated dendrogram) on the dendrogram.
- **Colors**: Activate this option to use colors to represent the different groups on the full dendrogram.

Display charts on two axes: Activate this option if you want the different graphical representations to be displayed only on the first two axes.

CA eigenvalues: Activate this option to display the *scree plot* of the CA eigenvalues in each class.

CA biplot: Activate this option to display the plot of the consensus coordinates of each class in the factor space.

Similarity assessors/consensus: Activate this option to display the bar chart of the similarity index between each assessor and the consensus of its class.

Weights: Activate this option to display the bar chart of the weights created and used by CATATIS.

Results

Descriptive statistics: The number of checks by assessor is displayed.

Similarity matrix (S): The matrix of similarity index between all assessors is displayed. The similarity index is included between 0 and 1. The closer it is to 1, the stronger the similarity. This

index is the Ochiai coefficient.

Node statistics: This table shows the data for the successive nodes in the dendrogram. The first node has an index which is the number of assessors increased by 1. Thus, it is easy to see at any time if an assessor or group of assessors is clustered with another group of assessors in the dendrogram.

Levels bar chart: This table displays the statistics for dendrogram nodes, which correspond to the increase in the CLUSCATA minimization criterion (equivalent to the increase in within-class variance) when merging two classes.

Dendrograms: The full dendrogram displays the progressive clustering of assessors. If truncation has been requested, a broken line marks the level the truncation has been carried out. The truncated dendrogram shows the classes after truncation.

Composition of classes:

Results by assessor: This table shows the assignment class for each assessor in the initial assessors order. If a consolidation is requested, the results are given before and after the consolidation. If you have checked "class K+1", it is possible that some assessor may have a missing value after consolidation. This means that they are not placed in any of the main classes (they are placed in class "K+1").

Results by class: The results are given by class. Thus, a list of assessors is displayed for each class.

Number of assessors per class: The number of assessors in each class is indicated.

Rho parameter computed: Result displayed only if you have chosen to add the class "K+1". The rho parameter represents the minimum similarity that an assessor must have with the consensus of a class in order to belong to it. If this condition is not met for any of the classes, the assessor is placed in class "K+1". This parameter is calculated according to the proximity of each assessor to its class as well as to the neighboring class.

Analysis of the class k:

In this section, the analysis of each of the classes by the CATATIS method is displayed.

Eigenvalues of CA: The eigenvalues of CA and corresponding chart (*scree plot*) are displayed.

Product coordinates: The coordinates of the products of the consensus in the factors space are displayed, with the corresponding charts (depending on the number of factors chosen).

Attribute coordinates: The coordinates of the attributes of the consensus in the factors space are displayed, with the corresponding charts (depending on the number of factors chosen).

Consensus configuration: The consensus configuration is displayed. It corresponds to the weighted average of the assessors data.

Similarity assessors/consensus: The similarity indices between the assessors and the consensus are displayed, with the associated bar chart. Like the weights of CATATIS, these coefficients allow to detect atypical assessors. The advantage of these coefficients is that they are between 0 and 1, so they are easier to interpret than the weights.

Weights: The weights calculated by CATATIS are displayed, with the associated bar chart. The greater the weight, the more the assessor contributed to the consensus. Knowing that CATATIS gives more weight to the closest assessor from a global point of view, a much lower weight than the others will mean that the assessor is atypical.

Indices:

Homogeneities: The homogeneity of each class is displayed. It is a value between $1/m$ (m being the number of assessors of the class) and 1, which increases with the homogeneity of the assessors. In a second step, the global homogeneity, which is a weighted average of the homogeneity of each class, is displayed.

Global Error/Within-class Variance: The error of the CLUSCATA criterion is displayed. It corresponds to the within-class variance.

Example

A tutorial on how to use CLUSCATA is available on the XLSTAT Help Center:

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

References

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).

Semantic differential charts

Use this method to easily visualize on a chart, ratings given to objects by a series of judges on a series of dimensions.

In this section:

Description

Psychologist Charles E. Osgood has developed the visualization method *Semantic differential* in order to plot the differences between individuals' connotations for a given word. When applying the method, Osgood asked survey participants to describe a word on a series of scales ranging from one extreme to the other (for example favorable/unfavorable). When patterns were significantly different form one individual to the other or from one group of individuals to the other, Osgood could then interpret the Semantic Differential as a mapping of the psychological or even behavioral distance between the individuals or groups.

This method can also be used for a variety of applications:

Analysis of the experts' perceptions for a product (for example a yogurt) described by a series of criteria (for example, acidity, saltiness, sweetness, softness) on similar scales (either from one extreme to the other, or on a similar likert scale for each criterion): a Semantic differential chart will allow to quickly see which experts agree, and if significantly different patterns are obtained.

Survey analysis after a customer satisfaction survey.

Profile analysis of candidates during a recruitment session.

This tool can also be used in sensory data analysis. Here are two examples:

A panel of experts rates (from 1 to 5) a chocolate bar (the object) on three criteria (the "attributes") namely the flavor, the texture, the odor. In this case, the input table contains in cell (i,j) the rating given by the i-th judge to the product on the j-th criterion. The semantic differential chart allows to quickly compare the judges.

A panel of experts rates (from 1 to 5) a series of chocolate bars (the objects) on three criteria (the "attributes") namely the flavor, the texture, the odor. In this case, the input table contains in cell (i,j) the average rating given by the judges to i-th the product on the j-th criterion. The semantic differential chart allows to quickly compare the objects.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

General tab:

Descriptors: Select the descriptor data from the Excel sheet. If the first line of the selection includes headings, the option "Labels of descriptors" should be enabled.

Objects: select the labels of the objects on the Excel sheet.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Descriptor labels: Activate this option if the first row of the selected data (data, observation labels) contains a header.

Charts tab:

Color: Activate this option to use different colors when displaying the lines corresponding to the various objects/descriptors.

Grid: Activate this option to display the grid on the chart.

Values: Activate this option to display the values on the chart.

Results

The result that is displayed is the Semantic Differential chart. As it is an Excel chart, you can modify it as much as you want.

Example

An example of Semantic Differential Charts on the XLSTAT Help Center:

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

References

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.

TURF Analysis

Use this tool to run a TURF (Total Unduplicated Reach and Frequency) analysis to highlight a group of products that will reach better market share.

In this section:

Description

The TURF (Total Unduplicated Reach and Frequency) method is used in marketing to highlight a line of products from a complete range of products in order to have the highest market share. From all the products of a brand, we can obtain a subset, which should be the line of products with the maximum reach.

For example, let's consider an ice cream manufacturer producing 30 different flavors and who wants to put forward a line of six flavors that will reach as many consumers as possible. Thus, he submitted a questionnaire to a panel of 500 consumers who scored each flavor on a scale from 1 to 10. The manufacturer believes that the consumer will be satisfied and inclined to choose the flavor if he gives a score above 8. TURF analysis will look for the combination of 6 flavors with greatest reach and frequency.

This method is a simple statistical method. It is based on a questionnaire (with scores on a fixed scale). The analysis runs through every possible combination of products and records for each combination (1) the percentage of those that desire at least 1 product in the given combination (i.e. reach), and (2) the total number of times products are desired in the given combination (i.e. frequency).

XLSTAT offers a variety of techniques to find the best combination of products: The enumeration method will test all the combinations but may be time consuming; the greedy algorithm is very fast but can stop on a local optimum and the fast search algorithm is close from the enumeration method but it is faster and does not guarantee the optimal solution.

Methods

The data used are data from a questionnaire: one row per consumer and one column per product. They should be in the form of scores (Likert scales). XLSTAT allows you to define different scales. However, all notes must be on the same scale. The user chooses an interval in which he considers that the goal is reached (eg scores greater than 8 in 10).

XLSTAT allows you to use three different algorithms to find the right product line:

- **The enumeration method**: All combinations of k products on the p products ($k < p$) are tested. We retain the combinations that have the highest reach and frequency.

The reach is defined by:

Reach = Number of consumers that desire at least 1 product in the given combination

This method is accurate but can be highly time consuming if p and k become large (eg for p = 40 and k = 12, there will be 5,586,853,480 combinations).

- **The greedy algorithm**: This algorithm is a simple heuristic that can find a good result very quickly by maximizing the reach.

It works as follows:

- Find the product hitting the target more often (highest frequency)

- Select this product in the combination

- Remove the observations for which this product has reached the target

- Repeat until there is no observation to remove or until you are not able to withdraw any observation

This algorithm is repeated many times with different initial conditions to minimize the risk of falling into local optima. Its advantage lies in its quickness but it does not warrant obtaining the global maximum.

- **The fast algorithm**: This algorithm is based on the same principle as the enumeration method but when no improvement is found after a number of steps, a jump in the combinations is done in order to avoid some "unnecessary" combinations. This algorithm does not guarantee to find the local optimum but it will explore more possibilities than the greedy algorithm and it is less costly than the enumeration method.

Constraints in TURF

XLSTAT allows to add constraints on the products in the framework of a TURF analysis. Two kinds of constraints are available:

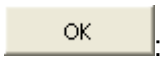
- Constraints on the products: In that case, you can force a product to be included in a line of products.

Constraints by group: In that case you select a supplementary variable associating a group to each product. When the algorithm runs, at least one product of each group will be included in

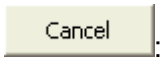
the line.

Dialog box

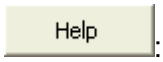
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data: Select the data in the form of scores on a common scale for all products. If headers have been selected, please check the option "Labels samples" is enabled.

Subset size: Select the size of the subset (that is to say the number of products that must be integrated into the product line).

Other groups: Select this option if you want to select another group of products in which a subset will also be obtained. Sometimes two categories of products have to be represented in the line. We want a certain number of products for each category. (In the case of the previous example, we may want three ice creams and three sherbets.)

Scale: Choose the scale used to rate products. If you select "other", then you must enter a minimum and the maximum of your scale.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selections (data, other group) contains a label.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection.

Observation weight s: Activate this option if observations weights are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection.

Constraints on the products: Activate this option if you want to include some products in all the line of products. If this option is activated, when you press Ok, a new dialog box will enable you to choose which product to include in your lines.

Constraints by group: Activate this option if you want that at least one of the products of each group is included in the generated line of products. Select a column in which the group associated to each product is given. This column should have as many elements as the number of products. If the "Variable labels" option is activated you need to include a header in the selection.

Options tab:

Number of combinations displayed: Enter the number of combinations you want to keep. It is sometimes interesting to look at several combinations that get good reaches and frequencies to select the best product line.

The objective is attained for scores between __ and __: Enter the lower and upper bounds of the scores that will be considered as reaching the objective.

Method: Select the method you want to use for analysis TURF. For the list, you can enter a maximum time for the algorithm stops automatically. If the number of combinations is reduced, XLSTAT automatically opts for the enumeration method because it is accurate.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations that contain missing data.

Results

Frequencies by product: This table displays the frequency with which the objective has been reached for each product.

Product lines obtained with the TURF analysis: This table displays for each selected combination: the Reach, the frequency and the name of each product.

Product lines obtained with the TURF analysis (%): This table displays for each selected combination: the percentage of observations for which the objective has been reached, the frequency in percentage, and the frequency in percentage for each product in each combination.

Two-way table (TURF): This table crosses all the products (rows) and all the line of products (columns). The first and second rows are the reach and reach in percentage for each line. The last column displays the frequency (in percentage) that each product has been included in the generated lines of products. When a product is included in a line, the value in the table is the global frequency divided by the reach (multiplied by 100).

No intetion to purchase: This result allows us to know the number of subject(s) with no purchase intent.

Example

An example of TURF analysis is available on XLSTAT Help Center:

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

References

Miaoulis G., Free V. and Parsons H. (1990). TURF: A new planning approach for product line extensions. *Marketing Research*, II (March), 28-40.

Krieger A. M. and Green P. E. (2000). TURF revisited: enhancements to total unduplicated reach and frequency analysis. *Marketing Research*, **12** (Winter), 30-36.

Sensory wheel

Use the sensory wheel tool to create a chart that displays the words used to describe a product. This tool allows you to create and display on the same chart a hierachy or to use it, if it is already available.

In this section:

Description

The sensory wheel tool allows to display on synthetic diagram (donought chart), a list of words or a classification of words used to describe a product. This kind of information could also be presented on a tree disagram, however a split on a tree means a clear alternative, which is not the case here. For example, on a sensory wheel, you can decide that the left part concerns the taste, and the right part the smell, without carrying any idea that pollutes the underanding of the chart: words displayed on the left part concern the taste and words on the right part concern the smell.

Ideally the classification should idealy be conceived such that a given word is present only once on the diagram.

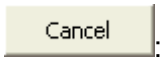
XLSTAT allows to use as input two different types of datasets:

- A classification table that describes a hierarchy between words,
- A list of unique words, with if available the frequencies or weights of the different words (by default, relative weights are considered equal). If some words are repeated in the list, repetitions can then either be considered or not.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data: Select the data whether they correspond to a **list of words** or a **classification table**. If headers have been selected, please check the option "Labels samples" is enabled.

Frequencies: Activate this option if you want to weight the words and take into account the weighting on the sensory wheel.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Labels included: Check this option if the first line of the selections (data, frequencies) contains a label.

Options tab:

Reformat words: Activate this option so that XLSTAT, on one hand identifies the repeated words and on another removes spaces that might be before or after each word.

Size with frenquencies: Activate this option so that XLSTAT adapts the size of the donoughts slices depending on the frequency of each word.

Number of levels: Enter the number of levels that should be taken into account (for nested wheels).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations that contain missing data.

Outputs tab:

Word frequencies: Activate this option so that XLSTAT displayes the frequencies of % of each word.

Results

If the option "option" has been activated, the first table displays the different words with for each word its frequency and the corresponding %.

The sensory wheel is then displayed. The button that precedes the wheel allows to activate or deactivate the chart. When it is active, a simple click on the chart displays a dialog box that makes possible five different actions:

- Merge two words
- Move one word after the other
- Rename a word
- Align the words with the radius
- Align the orientation of words perpendicularly to the radius

Example

An example showing how to generate a sensory wheel is available at the XLSTAT Help Center:

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

References

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.

Design of experiments for sensory data analysis

Use this tool to create an experimental design in the context of sensory data analysis.

In this section:

Description

Designing an experiment is a fundamental step for anyone who wants to ensure that data collected will be statistically usable in the best possible way. No use to evaluate products from a panel of judges if the products cannot be compared under statistically reliable conditions. It is also not necessary to have each judge evaluate all products to compare products between them.

This tool is designed to provide specialists in sensory analysis to provide a simple and powerful tool to prepare a sensory evaluation where judges (experts and/or consumers) evaluate a set of products.

When you want a panel of consumers to evaluate a set of products, say 9, the first issue that arises is what is the appropriate number of consumers that should be involved, knowing that there may be technical constraints (a limited number of trained consumers is available), or budgetary constraints. Once the number of consumer is defined, for example 82, arises the question of the maximum number of products that a consumer can evaluate during each session. Rarely for budgetary reasons, but mostly because of physiological constraints: a consumer, even trained, may not necessarily retain its sensory capabilities to rate too many products at once. Imagine that the experiment shows that three products is a maximum for a session and for organizational reasons, only two sessions can be arranged. Each consumer can then evaluate a maximum of 6 products.

It remains to determine which products will be evaluated by each of the 82 consumers in each session, and in what order. It is possible that the order has an influence (this is not the issue here, but a proper design of experiments would allow verifying or invalidating this assumption). To avoid penalizing certain products we should ensure that products are seen as often as possible in the three different positions during each session. Furthermore, it is possible that some sequences of products also have a bearing on sensory assessments. We restrict here to consider pairs of products (carry-over of order 2). As for the order, we will also ensure that different ordered pairs, 72 in our example, be present at a frequency as homogeneous as possible in the design.

When generating the plan we therefore try to reconcile the following three requirements:

Products must be seen by as many judges as possible and with an overall frequency of the different products as homogeneous as possible,

Each product must be seen in the different orders during each session, with an overall frequency for each pair (order, product) as homogeneous as possible

The different ordered pairs of products must be present in the design of experiments with a frequency as homogeneous as possible.

Measuring the performance of the design

Let N be the matrix with rows for products and columns for judges, and containing the frequency in which each judge sees each product during each session. In the sensory designs we are dealing with sensory, N contains either 0 or 1's (we impose that a jugde can evaluate only once a given product during a session) and the marginal sums for columns are constant and equal to k (the same number of products is being evaluated by each judge).

Maxtrix $M = NN'$ contains on its diagonal the frequency in the design of each product, and on the triangular parts the number of times each unordered pair of products has been evaluated by a same judge. This matrix is called the **concurrence matrix**.

Matrix $A^* = I - qNN'qk$ (q is a diagonal matrix with the inverse of the square root of the frequency of each product in the design) is directly related to the products information matrix, and as as consequence, to the variances and covariances of the parameters of the products in the ANOVA model we can compute once the ratings have been recorded. If we want to ensure that the variances of the differences between the parameters associated with the products are as uniform as possible, we need to ensure that the eigenvalues of the A^* matrix are close to each other.

We define the **A-efficiency** as the harmonic mean of the at most $p - 1$ nonzero eigenvalues of matrix A^*, and the **D-efficiency** as the geometric mean of the same values. The two criteria are equal in the ideal case where all eigenvalues are equal.

Balanced Incomplete Block Designs

A block design is a design in which we study the influence of two factors on one or more phenomena. We know that one factor has an impact that we cannot control, but that is not of interest. So we want to ensure that this factor does not disturb the analysis that we perform once the collected data. For this we make sure that the various levels of other factors are well represented in each block.

In our case, the blocking factor corresponds to the judges, and the factor of interest we want to study corresponds to the products.

A complete block design is a design in which all levels of the factors of interest are present once within each block. For a sensory design, this corresponds to a design where all products are

seen once by each judge.

In an incomplete block design, all levels of the factors of interest are not present for all levels of the blocking factor. It is balanced if each level of the factor of interest are present a same number of times r in the design, and if each pair of levels of each factor is present the same number of times λ .

If v is the number of products studied, b the number of judges, k the number of products seen by each judge, we show that the following conditions are necessary (but not sufficient) to have a balanced incomplete block design:

$$bk = vr$$

$$r(k-1) = \lambda(v-1)$$

In the case where a balanced incomplete block design exists, the optimal value of the two criteria (A and D-efficiency) is known. We have:

$$E = \frac{\nu(k-1)}{k(\nu-1)}$$

XLSTAT allows users to search an optimal design within the meaning of the A-efficiency or the D-efficiency, and whether in the case of complete plans or in the case of incomplete block designs, whether balanced or not.

Algorithm to obtain a design

XLSTAT relies on two different techniques for generating designs. If the "Fast" option is selected by the user and if $(\frac{b}{v})$ is integer, then XLSTAT uses the cyclic plans initiated by Williams (1949), and studied in great detail by John and Williams (1995). If (b/v) is not an integer or if the "Fast" option has not been required, XLSTAT uses a proprietary (unpublished) algorithm to very quickly generate an appropriate solution and, starting from this solution, it searches by simulated annealing a better solution for a time defined by the user. If the design is a complete or balanced incomplete block design and if the optimum is found, the search is interrupted before the available time runs out.

Algorithms to improve the columnfrequency or the carry-over

Once the design is found (the matrix N is known), we need to order products to optimize the in terms of column frequency and carry-over (Périnel and Pagès, 2004). We want that each product is present the same number of times at a given position, and that each ordered pair is also present the same number of times. In order to obtain that, XLSTAT uses two matrices: the matrix of column frequencies and the matrix of carry-over. The goal of algorithm is to make as homogeneous as possible these matrices. The matrix of the carry-over is a matrix that shows the position ij the number of times that product i precedes product j in the design. We define a parameter lambda which will allow us to favor either obtaining a good carry-over (λ close to 0) or obtaining a column frequency matrix positions closest to the constant matrix (λ close to 1).

The optimization algorithm is iterative. It permutes the ranks of the products for each judge to maximize the following criterion:

$$\lambda \sum_{i,j} r_{ij}^2 + (1-\lambda) \sum_{i,j} s_{ij}^2$$

where r_{ij} are the elements of the column frequency matrix and s_{ij} are the elements of the carry-over matrix. As soon as an optimal value is reached or as the maximum number of iterations is reached the algorithm stops.

XLSTAT uses two criteria to verify the quality of the two matrices:

MDR (mean deviation of R): $MDR = \sum_{i,j}(r_{ij} - \overline{r})$, which is the deviation for the elements of the column frequency matrix.

MDS (mean deviation of S): $MDS = \sum_{i,j}(s_{ij} - \overline{s})$, which is the which is the deviation for the elements of the carry-over matrix

For incomplete block designs, one can calculate the optimal value that should be reached.

Resolvable designs and improved presentation

A design is said to be resolvable if it can be divided into g groups of judges that are such that, within each group, we have a unique instance of each product. Some balanced incomplete block designs have this property. Presenting a design using such a subdivision into groups has the advantage that if some judges do not appear, it is not necessary to rebuild a design of experiment, but simply to ensure that the last experiments are canceled. This approach is also particularly interesting when one wants to put in place several evaluation sessions (see below). A condition for a balanced incomplete block design to be resolvable is that $\frac{v}{k}$ must be an integer

Even when they are balanced or not resolvable, XLSTAT tries to present the incomplete block designs so that the products are present at most twice and, if possible once in a group of size $< \frac{v}{k} >$ (where $< i >$ equals i if i is an integer, and the rounding to the next integer value otherwise). Thus, if judges were finally absent, it does not penalize too much the quality of the design.

Sessions

It is sometimes necessary to split sensory evaluations into sessions. To generate a design that takes into account the need for sessions, XLSTAT uses the same intial design for each session and then applies permutations to both rows and columns, while trying to keep as even as possible column frequencies and carry-over. When the designs are resolvable or near resolvable, the same judge will not be testing twice the same product during two different sessions.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

General tab:

Product: Enter the number of products involved in the experiment.

Products/Judge: Enter the number of products that should evaluate each judge. If the session option is activated, you need to enter the number or products evaluated by each judge during each session.

Judges: Enter the number of judges evaluating the products.

Sessions: Activate this option if the design should comprise more than one tasting session.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Judge labels: Activate this option if you want to select on an Excel sheet the labels that should be used for the judges when displaying the results.

Options tab:

Method: Choose the method to use to generate the design.

- **Fast**: Activate this option to use a method that reduces as much as possible the time spent to find a fine design.

- **Search**: Activate this option to define the time allocated to the search for an optimal design. The maximum time must be entered in seconds.

Crite rion: Choose the criterion to maximise when searching for the optimal design.

- **A-efficiency**: Activate this option to search for a design that maximizes the A-efficiency.
- **D-efficiency**: Activate this option to search for a design that maximizes the D-efficiency.

Carry-over vs order effect: Define here your preference regarding what should be the priority in the second phase when generating the design: choose among homogenizing the **order effect** which is the frequency of the products rankings (the order in which they are evaluated), or homogenizing the number of times two products are evaluated one after the other (**carry-over**).

- **Lambda**: Let this parameter vary between 0 (priority given to carry-over) and 1 (priority given to column frequency).
- **Ite rations**: Enter the maximum number of iterations that can be used for the algorithm that searches for the best solutions.

Product codes: Select how the product codes should be generated.

- **Product ID**: Activate this option to use a simple product identifier (P1,P2, …).
- **Random code**: Activate this option to use a random three letters code generated by XLSTAT.
- **User defined**: Activate this option to select on an Excel sheet the product codes you want to use. The number of codes you select must correspond to the number of products.

Outputs tab:

Judges x Produc ts table: Activate this option to display the binary table that shows if a judge rated (value 1) or not (value 0) a product.

Concurrence table: Activate this option to display the concurrence that shows how many times two products have been rated by the same judge.

Judges x Ranks table: Activate this option to display the table that shows, for each judge, which product is being rated at each step of the experiment.

Order effect table: Activate this option to display the table that shows how many times each product has been rated at a given step of the experiment.

Carry-over table: Activate this option to display the table that shows how many times each product has been rated just after another one.

Design table: Activate this option to display the table that can later be used for an ANOVA, once the ratings given by the judges have been recorded.

Results

Once the calculations are completed, XLSTAT indicates the time spent looking for the optimal plan. The two criteria A and D-efficiency are displayed. XLSTAT indicates if the optimal plan has been found (case of a balanced incomplete block design). Similarly, if the plan is resolvable, it is indicated and the group size is specified.

If sessions have been requested, a first set of results is displayed with results taking into account all the sessions. The results for each session are then displayed

The **Judges x Products table** is displayed to show whether a judge has assessed (value 1) or not (value 0) a product

The **concurrence table**: shows how many times two products have been rated by the same judge.

The **MDS/MDR table** displays the criteria that allow assessing the quality of the column frequencies and carry-over that have been obtained, compared to the optimal values.

The **Judges x Ranks table** shows, for each judge, which product is being rated at each step of the experiment.

The **order effect table** shows how many times each product has been rated at a given step of the experiment.

The **carry-over table** shows how many times each product has been rated just after another one.

The **design table** can later be used for an ANOVA, once the ratings given by the judges have been recorded.

Example

An example showing how to generate a DOE for sensory data analysis is available on XLSTAT Help Center:

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

References

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.

Pé rinel 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.

Design experiments for sensory discrimination tests

Use this tool to create an experimental design in the context of sensory discrimination tests. This tool allows you to generate the setting for different tests: triangle, duo-trio, two-out-of-five, 2-AFC, 3-AFC and tetrad.

In this section:

Description

Designing an experiment is a fundamental step for anyone who wants to ensure that data collected will be statistically usable in the best possible way. There's no point in evaluating products from a panel of assessors if the products cannot be compared under statistically reliable conditions.

This tool is designed to provide specialists in sensory analysis a simple and powerful tool to prepare a sensory discrimination test where assessors (experts and/or consumers) evaluate a set of samples.

Before introducing a new product on the market, discrimination testing is an important step. XLSTAT allows you to prepare the tests. With XLSTAT, you can generate a combination of products to be presented to your assessors so that they are in the correct setting for that kind of test.

Sensory discrimination tests are based on comparing two products that are presented in a specific setting.

When creating your design, you have to know which test you want to apply, the number of assessors and, if possible, the products' names.

XLSTAT allows you to run these tests:

- Triangle test: 3 products are presented to each assessor in different orders. Within these products, 2 are similar and the third one is different. Assessors have to identify the product that is different from the others.
- Duo-trio test: Assessors taste a reference product. Then they taste two different products. Assessors must identify the product that is similar to the reference product.

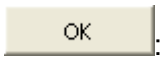
- Two-out-of-five test: 5 products are presented to the assessors. These products are separated into two groups, the first one with 3 identical products and the second one with 2 identical products. The assessors have to identify the group with 2 identical products.
- 2-AFC test: 2 products are presented to each assessor. The assessors have to tell which product has the highest intensity for a particular characteristic.
- 3-AFC test: 3 products are presented to each assessor. Two are similar and the third one is different. The assessors have to tell which product has the highest intensity on a particular characteristic.
- Tetrad test: 4 products are grouped into two groups, with identical products within each group are presented to each assessor. The assessors are asked to distinguish between the two groups.

For each test, you can generate experiment designs using randomization of the available combinations.

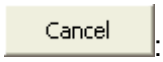
You can specify more than one session and add labels to the assessors and products.

Dialog box

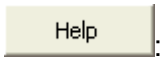
The dialog box is divided into several tabs that correspond to a variety of options, ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help section.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.


: Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Type of test: Select the name of the discrimination test you want to use.

Judges: Enter the number of judges evaluating the products.

Sessions: Activate this option if the design should comprise more than one tasting session.

Judge labels: Activate this option if you want to select on an Excel sheet the labels that should be used for the judges when displaying the results.

Column label: Activate this option if column headers have been selected within the selections.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Product codes: Select how the product codes should be generated.

- **Product ID**: Activate this option to use a simple product identifier (P1,P2, etc.).
- **Random code**: Activate this option to use a random code generated by XLSTAT. Two options are available: the first, *Alphabetical*, allows you to generate a random code with three letters, and the second, *Numeric*, allows you to generate a random code with three numbers.
- **User defined**: Activate this option to select on an Excel sheet the product codes you want to use. Two columns are required (one for each product), and the number of selected rows must correspond to the number of different samples required for the experiment design. Thus, 2 rows are required for the triangle, Duo-Trio, 3-AFC and Tetrad tests. The 2-AFC test only requires one row and the Two-out-of-Five test requires 3 rows. Finally, it should be noted that in the case of the Duo-Trio and 3-AFC tests, the cell (2,2) of the selected data must be empty since product 2 only requires one sample.

Results

Once the calculations are completed, XLSTAT displays the question to be asked to the assessors specific to the chosen test.

The next table displays the product that should be tasted by each assessor (one row = one assessor, one column = one sample). The last column is left empty to allow you to enter the result of the tasting. Note that for the Triangle, Duo-Trio, 2-AFC and 3-AFC tests, the correct/incorrect answer will be automatically filled if the user enters the answers of each judge into the design table. For 2-AFC, you will have to enter the correct answer (which is the same for the whole column) to use this feature.

Example

An example showing how to generate a DOE for a discrimination test together with the analysis of the results is available on XLSTAT Help Center:

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

References

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.

Sensory discrimination tests

Use this tool to perform discrimination tests, among which the triangle test, the tetrad test, the duo-trio test, the 2-AFC test, the 3-AFC test or the two-out-of-five test.

In this section:

Description

Before introducing a new product on the market, discrimination testing is an important step. XLSTAT allows you to prepare the tests (see design of experiments for discrimination tests) and to analyze the results of these tests.

Three models can be used to estimate the parameters of these tests:

- The guessing model.
- The Thurstonian model.
- The Beta-Binomial model.

XLSTAT allows you to run:

- Triangle test: 3 products are presented to each assessor in different orders. Within these products, two are similar and the third one is different. Assessors must identify the product that is different from the others.
- Duo-trio test: Assessors taste a reference product. Then they taste two different products. Assessors must identify the product that is similar to the reference product.
- Two-out-of-five test: 5 products are presented to the assessors. These products are separated into two groups, the first one with 3 identical products and the second one with 2 identical products. The assessors must identify the group with 2 identical products.
- 2-AFC test: 2 products are presented to each assessor. The assessors must tell which product has the highest intensity for a particular characteristic.
- 3-AFC test: 3 samples are presented to each assessor. Two are similar and the third one is different. The assessors must tell which product has the highest intensity on a particular

characteristic.

- Tetrad test: 4 products grouped into two groups, with identical products within each group are presented to each assessor. The assessors are asked to distinguish the two groups.

Each of these tests has its own advantages and drawbacks. A complete review on the subject if available in the book by Bi (2008).

Some concepts should be introduced: pC is the probability of a correct answer, pD is a probability of discrimination, pG is the guessing probability, d' is the d-prime also called Thurstonian delta. These concepts are detailed below.

Models

Two models are commonly used in discrimination testing:

The guessing model assumes that consumers are either discriminators or non-discriminators. Discriminators always find the correct answer. Non-discriminators are guessing the answer with a known guessing probability (which depends on the test used). For example, someone who does not taste a difference will still have 1 chance out of 3 for the triangle test. The proportion of discriminators is the proportion of people who are able to actually detect a difference between the products.

This concept can be expressed as $pD = \frac{(pC - pG)}{(1 - pG)}$ where pC is the probability of a correct answer and pG is the guessing probability.

In the Thurstonian model, the required parameter is not a probability of discrimination pD but a d' (d-prime). It is the sensory distance between the two products, where one unit represents a standard deviation.

The assumptions are that the sensory representations of the products are following two normal distributions and that the consumers are not categorized as discriminators/non-discriminators. Consumers are always correct, translating what they perceive. Thus, an incorrect answer is translated into closeness between products that leads to an incorrect perception. If d' is close to 0, then products cannot be discriminated.

For each test, you will have the guessing probability (as in the guessing model) and a psychometric function that link d' and the probability of correct answers. These parameters are specific to each test.

We have $pC = f_{\text{test}}(d')$

Guessing probability

For each test the guessing probability which is the probability to obtain the correct answer by guessing is equal to:

Triangle test: $pG = 1/3$

Duo-trio test: $pG = 1/2$

Two-out-of-five test: $pG = 1/10$

2-AFC: $pG = 1/2$

3-AFC: $pG = 1/3$

Tetrad test: $pG = 1/3$

Psychometric functions

For each test the psychometric function which is the link between d' and pC (the probability of a correct answer) is defined by:

Triangle test: $pC = f_{triangle}(d^{'}) = 2\int_0^{\infty}\{\Phi[-x\sqrt{3} + d^{'}\sqrt{2/3}] + \Phi[-x\sqrt{3} - d^{'}\sqrt{2/3}]\}\phi(x)dx$

Duo-trio test: $pC = f_{duo-trio}(d^{'}) = -\Phi(d'/\sqrt{2}) - \Phi(d'/\sqrt{6}) + \Phi(d'/\sqrt{2})\Phi(d'/\sqrt{6})$

2-AFC: $pC = f_{2-AFC}(d^{'}) = \Phi(d^{'}/\sqrt{2})$

3-AFC: $pC = f_{3-AFC}(d^{'}) = \int_{-\infty}^{\infty}\phi(x - d^{'})\Phi[x]^2 dx$

Tetrad test: $pC = f_{\text{tetrad}}(d^{'}) = \int_{-\infty}^{\infty}\phi(x)\Phi[x]\{1 - \Phi[x - d^{'}]\}^2 dx$

These functions are estimated using the Gauss-Kronrod algorithm for numerical integration.

Calculating p-value and power

p-value and power for these tests are obtained using the binomial or normal distribution based on the estimated pC. XLSTAT allows the possibility of carrying out either a difference test or a similarity test. And in the case of the 2-AFC test, XLSTAT gives the possibility of performing a unilateral or bilateral test.

Standard error and confidence intervals for the Thurstonian model parameters

When using the Thurstonian model, you can obtain standard error and confidence interval for the parameters of interest.

For the probability of a correct answer pC, we have:

$$SE(pC) = \sqrt{pC(1 - pC)/N}$$

where N is the number of assessors.

For the probability of discrimination pD, we have:

$$SE(pD) = \frac{SE(pC)}{1 - pG}$$

For the d', we have:

$$SE(\hat{d}) = \frac{SE(pC)}{f'_{test}(\hat{d})}$$

Where f' is the derivative of the psychometric function with respect to d' (Brockhoff and Christensen, 2010).

Beta-Binomial model

The binomial model, used for the guessing model and the Thurstonian model, is based on the assumptions that the choices of each assessor are independent and that the probability of a correct answer is the same for all. However, when sessions are introduced into the data, these assumptions are no longer respected. We talk about sessions when a sensory discrimination test performed on k assessors is repeated on all or part of the assessors until n times. In this case, the independence between the answers is no longer respected because they can be given by the same assessors, and thus the probability of choosing the correct answer varies according to the assessors' discrimination ability. This is called overdispersion because the data contains several sources of variation.

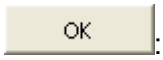
The Beta-Binomial model allows to consider this phenomenon in the test results by estimating the parameters μ and $gamma$. The estimate of μ represents the probability of a correct answer pC. The parameter $gamma$ measures the variation between assessors. If $gamma$ is close to 0, it means that there is no overdispersion, which is the same as using the binomial model. On the contrary, if $gamma$ is close to 1, the Beta-Binomial model must be used to respect the underlying assumptions, and avoid underestimation of the standard deviations and misleading interpretation.

The parameters of the Beta-Binomial model are estimated by maximum likelihood.

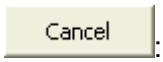
Through the Beta-Binomial model, we will test both if the probabilities of a correct answer from each assessor are equal to the guessing probability pG, as well as the existence of a variation between the assessors. If one of the two test hypotheses is rejected ($\mu = pG$ or $gamma = 0$), the products are considered different by the assessors.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Type of test: Select the test you want to analyze.

Method: Select the model to be used between the Thurstonian model and the guessing model. To use the Beta-Binomial model, select *Data with sessions* in drop-down list *Input data*.

Input data: Select the type of data you want to select as input. Four options are available depending on the chosen option, other options will be displayed.

- **Data selection case:**

Test results: Select a column with as many rows as assessors in which each cell gives the result of the test for each assessor.

Code for correct answer: Enter the code used to identify a correct answer in the selected data.

- **Sample size case:**

Number of assessors: Enter the total number of assessors in the study.

Number of correct answers: Enter the number of assessors that gave a correct answer to the test.

- **Proportion case:**

Number of assessors: Enter the total number of assessors in the study.

Proportion of correct answers: Enter the proportion of assessors that gave a correct answer to the test.

- **Data with sessions selection case:**

Table with sessions: Select a table with 2 columns. The first column contains the number of correct answers given for each assessor. The second column contains the number of sessions in which each assessor attended.

The following options will appear only if the Thurstone model is selected. They allow to specify the hypotheses of the test that must be verified.

Null hypothesis:

D-prime: Activate this option if you want to enter a fixed value for d'. You can then enter the value in the available textbox. The null hypothesis of the test is then "d' *is equal to* x", with x the chosen value.

pD: Activate this option if you want to enter a fixed value for the probability of discrimination. You can then enter the value in the available textbox. The null hypothesis of the test is then "*The probability of discrimination is equal to* x", with x the chosen value.

Note: by default, the test hypothesis that is tested is if d' is equal to 0.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the selected data contains a label.

Significance level (%): Enter the significance level for the test (default value: 5%).

Statistic: Select the method to use to compute the confidence intervals.

Power: Select the distribution to be used to compute the power.

Note: in the case of the Beta-Binomial model, the confidence intervals are calculated using the Wald method.

Test: Select the type of test you want to perform. Either a difference test or a similarity test.

In the case of a 2-AFC test, in addition to the difference test and the similarity test, it is also possible to choose whether you wish to carry out a bilateral test.

Results

Summary of selected options: This table displays the parameters selected in the dialog box.

Minimum (maximum) number of correct answers for the test to be significant: this table displays the minimum number (or maximum in the case of a similarity test) for the test to be significant.

Sensory discrimination tests: This table displays the results of the test performed, starting with the estimated value of the parameter of interest, d-prime or pD for the Thurstone model,

and μ (mu) and $gamma$ for the Beta-Binomial model. The interpretation of the test, as well as a p-value and a power are indicated.

Estimated parameter: This table displays the probabilities, the d-prime, the parameters μ and $gamma$ of the Beta-Binomial model if it has been selected, as well as their standard deviations and the associated confidence intervals. This table is not displayed in the case of the guessing model.

Example

A first example of discrimination test in sensory analysis is available on the XLSTAT Help Center at:

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

and another example with data including sessions is available at:

https://help.xlstat.com/6411-sensory-discrimination-test-sessions-excel

References

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.

Power - Sensory discrimination tests

This tool allows you to control the power or number of subjects in sensory discrimination tests. The discrimination tests available in XLSTAT are the triangle, duo-trio, two-of-five, 2-AFC, 3-AFC and tetrads tests.

Using this tool, you will be able to:

- determine the power according to the number of subjects
- determine the number of subjects according to the power

In this section:

Description

Experimental planning is a fundamental step for anyone who wants to ensure that the collected data will be usable under the best possible statistical conditions. There is no point in having products evaluated by a panel of subjects if they cannot then be compared under satisfactory statistical conditions.

XLSTAT Power for Sensory discrimination tests aims to provide sensory analysis specialists with a simple and powerful tool prior to the implementation of a sensory discrimination test in order to evaluate a set of products.

When testing a hypothesis using a statistical test, there are several decisions to take:

- The null hypothesis H0 and the alternative hypothesis Ha.
- The statistical test to use.
- The type I error also known as alpha. It occurs when one rejects the null hypothesis when it is true. It is set a priori for each test and is 5%.

The type II error or beta is less studied but is of great importance. In fact, it represents the probability that one does not reject the null hypothesis when it is false. We cannot fix it upfront, but based on other parameters of the model we can try to minimize it. The power of a test is calculated as $1 - beta$ and represents the probability that we reject the null hypothesis when it is false.

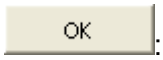
We therefore wish to maximize the power of the test. XLSTAT computes the power when other parameters are known. For a given power, it also allows to calculate the sample size (number of

subjects) that is necessary to reach that power.

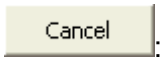
For more details on sensory discrimination tests, please refer to the Designs for sensory discrimination tests as well as to the Sensory discrimination tests.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Goal: Choose between computing power and sample size estimation.

test: Select the test you want to apply.

Alpha: Enter the value of the type I error.

Power (when sample size estimation has been selected): Enter the value of the power to be reached.

Sample size (when power computation has been selected): Enter the sample size (number of subjects).

Null hypothesis:

D-prime: Activate this option if you want to use a fixed value for d'. Then enter the value in the available textbox. The null hypothesis of the test is then "d' *is equal to* x", with x the chosen value.

pD: Activate this option if you want to use a fixed value for the probabibilty of discrimination. Then enter the value in the available textbox. The null hypothesis of the test is then "*The probability of discrimination is equal to* x", with x the chosen value.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select any cell in your Excel sheet.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Power: Select the distribution to be used to computer the power.

Results

A table of results is displayed. It is composed of 2 columns: the proportion of correct answers, followed by the sample size (or power depending on the parameters selected in the dialog box). It is used to construct the simulation graph which displayed below the table.

Example

A tutorial on how to determine the power or number of subjects in sensory discrimination tests is available on XLSTAT Help Center:

https://www.xlstat.com/demo/trp_en

References

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.

Create a Products/Assessors table

Use this tool to transform your vertical sensory data into a Products/Assessors table (horizontal).

In this section:

Description

Sensory data can be presented in two ways:

1. **Vertical form**: This is a table with **as many rows as samples**. One column is dedicated to the product identifier, another to the assessor identifier, and the remaining columns correspond to descriptors such as sweetness, acidity, etc. The size of the table is determined by the formula *n \times p + 2*, where n represents the number of samples and p represents the number of descriptive variables.

2. **Horizontal form**: This is a table with **as many rows as products**. The number of columns is equal to the product of the number of assessors and the number of descriptors. The size of the table is therefore *n \times (p \times m)*, where n represents the number of products, p represents the number of descriptive variables, and m represents the number of assessors. In the horizontal form, the data for the first assessor is displayed in the first p columns, the data for the second assessor is displayed in the next p columns, and so on.

A **Products/Assessors table**, representing the sensory data in a horizontal form, is required for certain sensory analyses in XLSTAT: * Multiple Factor Analysis (MFA). * Generalized Procrustes Analysis (GPA). * STATIS. * CLUSTATIS. * Projective mapping data analysis.

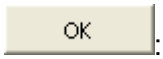
However, the sensory data may only be available in a vertical form. This functionality allows you to convert vertical sensory data into a **Products/Assessors table**.

Session Management

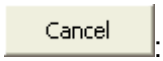
The **Create a Products/Assessors table** functionality automatically manages sessions (or repetitions) in cases where the same assessor tastes or tests the same product multiple times. When a product/assessor combination is identified multiple times, the average data for each descriptor is calculated.

Dialog box

You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files  .

Data: Select the columns corresponding to the descriptors. If column headers have been selected, please make sure the "Variable Labels" option is activated.

Products: Select the column containing the product identifiers. If column headers have been selected, please make sure the "Variable Labels" option is activated.

Assessors: Select the column containing assessor identifiers. If column headers have been selected, please make sure the "Variable Labels" option is activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections contains a label.

Show report header: Deactivate this option if you want the results table to start from the first row of the Excel worksheet (in the case of output to a worksheet or workbook), rather than after the report header.

Show assessor labels: Enable this option to display assessor labels at the top of the Products/Assessors table.

Results

By activating the **Show report header** option, several descriptive pieces of information are displayed: * The different selections for data, products, and assessors. * The number of identified assessors. * The number of identified products. * The number of sessions.

The **Products/Assessors table** is displayed and includes the following elements: * The first row displays the labels of the **assessors** (if the option is activated). * The second row displays the labels of the descriptors, which correspond to the same labels entered in the **Data** field. * The first column contains the labels of the **products**. * Finally, the body of the table presents the different data sorted by assessors.

Example

A tutorial on creating a Products/Assessors table is available in the XLSTAT Help Center:

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

JAR multivariate analysis and clustering

Use this tool to perform multivariate analysis (CATATIS) or clustering (CLUSCATA) on JAR (*Just About Right*) data.

In this section:

Description

JAR data is often only processed with penalty analysis, but they are packed with other information. In fact, just like the other tests, they can be used to describe products, to see similarities and differences, and so on. Thus, performing a multivariate analysis, at the same time creating a map of products with their descriptions, is very instructive.

JAR data

JAR data is data measured on a JAR (*Just About Right*) scale of 5, 7 or 9 levels. In the case of a 5 points JAR scale, 1 corresponds to « Not enough at all », 2 to « Not enough », 3 to « JAR » (*Just About Right*), an ideal for the consumer, 4 to « Too much » and 5 to « Far too much ». For example, for a chocolate bar, one can rate the bitterness, and for the comfort of the car, the sound volume of the engine.

CATATIS and CLUSCATA analysis of JAR data

By analyzing JAR data with CATATIS and CLUSCATA, it is then possible to (Llobell, 2022):

- Study links between products and attributes.
- Analyze the homogeneity of responses, which is highly informative about the quality of your data.
- Automatically construct groups of subjects with different points of view, thanks to an enhancement to the CLUSCATA method.

To use sensory analyses such as CATATIS and CLUSCATA on JAR data, you first need to pre-process your data. XLSTAT allows you to run a CATATIS or CLUSCATA analysis on JAR data without having to pre-process your data first.

Salton cosine

The agreement between two subjects is calculated using the Salton cosine which is equivalent to the Ochiai index (Salton & McGill,1983) in the case of binary data (Llobell, 2022):

$$s(X,Y) = \frac{< X,Y >}{||X||||Y||} = \frac{trace(XY^T)}{\sqrt{(trace(XX^T)trace(YY^T))}},$$

where 0 corresponds to complete disagreement and 1 to perfect agreement.

Pre-processing JAR data

To do this, the tool first transforms the JAR data into a horizontal table:

- The original data comprises one subject, three products and one attribute:

	Attribute 1
Product 1	Not enough
Product 2	JAR
Product 3	Too much

- A transformation from the table above to a disjunctive table is applied:

	Not enough	**JAR**	**Too much**
Product 1	**1**	0	0
Product 2	0	**1**	0
Product 3	0	0	**1**

- Finally, fuzzy coding is applied to the values for each product, to take account of the ordinality of the data:

	Not enough	**JAR**	**Too much**
Product 1	**1-**β	β	0
Product 2	$\beta/2$	**1-**β	$\beta/2$
Product 3	0	β	**1-**β

These steps are then applied to each attribute and then each fuzzy coding table is sorted and merged so that data from the first subject is displayed in the first p columns, data from the second subject is displayed in the next p columns, and so on, p being the number of attributes multiplied by three.

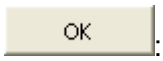
Here's an example of a table that can be obtained with two attributes, after this pre-processing:

	Assessor 1					
	Attribute 1			Attribute 2		
	Not enough	JAR	Too much	Not enough	JAR	Too much
Product 1	0.05	0.9	0.05	0.9	0.1	0
Product 2	0	0.1	0.9	0.05	0.9	0.05
Product 3	0.05	0.9	0.05	0	0.1	0.9

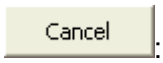
Finally, XLSTAT will use this table to perform a classic CATATIS or CLUSCATA analysis.

Dialog box

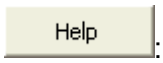
You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Just about right data: Select the data measured on the JAR scale. Several columns can be selected. If a column header has been selected, check that the "Column labels" option has been activated.

Scale: Select the scale that corresponds to the data (1 -> 5, 1 -> 7, 1 -> 9).

Products: Select the column containing the product identifiers. If column headers have been selected, please make sure the "Column Labels" option is activated.

Assessors: Select the column containing assessor identifiers. If column headers have been selected, please make sure the "Column Labels" option is activated.

*Note: it is important that **all subjects have seen all products** and that each **subject/product combination exists and exists only once**.*

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selections contains a label.

Analysis: Choose between one of the following two analyses:

- **CATATIS**: Choose this option to study and visualize links between products and attributes, and to study agreements between subjects using the CATATIS method.
- **CLUSCATA**: Choose this option to create homogeneous classes of subjects based on their perceptions of products, using the CLUSCATA method.

Options tab:

Beta: Specify the agreement parameter between JAR and other answers. Please enter a value between 0 and 0.5 (default value: 0.1).

Filter factors: You can activate one of the following two options to reduce the number of factors for which results are displayed.

- **Minimum %**: Enter the minimum percentage of the total variability that the chosen factors must represent.

- **Maximum Number**: Set the number of factors to consider.

CLUSCATA-specific options:

Truncation: Activate this option is you want XLSTAT to **automatically** define the truncation level, and therefore the number of classes to retain, or if you want to define the **number of classes** to create, or the **level** at which the dendrogram is to be truncated.

Consolidation: Activate this option to perform a consolidation of the classes obtained from the dendrogram.

Class K+1: Activate this option to add an additional class that will contain the assessors that do not fit any pattern of classes.

Rho parameter: Choose how you want to set the rho parameter: **automatically** or **user-defined**. This parameter represents the minimum agreement to be considered as sufficient in agreement to be kept in a class. The higher this parameter is set, the stronger the agreement required with the class and the more likely you are to place assessors in the K+1 class.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data has been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Products/Assessors table: Activate this option to display the Products/Assessors table built during the analysis.

CA eigenvalues: Activate this option to display the table and chart (scree plot) of eigenvalues of the CA on the consensus, or on the consensus of each class in the case of a CLUSCATA analysis.

CA coordinates: Activate this option to display the coordinates of the consensus in the factors space. In the case of a CLUSCATA analysis, the consensus coordinates of each class are displayed.

Similarity matrix (S): Activate this option to display the similarity matrix (Salton cosine).

Consensus configuration: Activate this option to display the consensus configuration or the consensus configuration of each class created by CATATIS.

Similarity assessors/consensus: Activate this option to display the similarity coefficient between each assessor and the consensus, or between each assessor and the consensus of its class in the case of a CLUSCATA analysis.

Weights: Activate this option to display the weights created and used by CATATIS.

Homogeneity: Activate this option to display homogeneity of the assessors and the homogeneity of each class in the case of a CLUSCATA analysis.

Global error: Activate this option to display the error of the CLUSCATA minimization criterion, equivalent to the intra-class variance, or to display the error of the CATATIS criterion.

CLUSCATA sub-tab:

Node statistics: Activate this option to display the statistics for dendrogram nodes.

Class composition: Activate this option to display the composition of each class.

Charts tab:

CA eigenvalues: Activate this option to display the *scree plot* of the CA eigenvalues, or the CA eigenvalues in each class in the case of a CLUSCATA analysis.

CA biplot: Activate this option to display the plot of the consensus coordinates in the factors space. In the case of a CLUSCATA analysis, the chart of the consensus coordinates of each class in the factor space will be displayed.

Display charts on two axes: Activate this option so that XLSTAT does not prompt you to select the axes, and automatically displays the biplot on the first two axes.

Weights: Activate this option to display the bar chart of the weights created and used by CATATIS.

Similarity assessors/consensus: Activate this option to display the bar chart of the similarity index between each assessor and the consensus. In the CLUSCATA analysis, these are the similarity coefficients between each subject and the consensus of its class.

CLUSCATA sub-tab:

Levels bar chart: Activate this option to display the diagram of levels showing the impact of successive clusterings on the within-class variance.

Dendrogram: Activate this option to display the dendrogram.

- **Horizontal**: Choose this option to display a horizontal dendrogram.
- **Vertical**: Choose this option to display a vertical dendrogram.
- **Truncated**: Activate this option to display the truncated dendrogram (the dendrogram starts at the level of the truncation).
- **Labels**: Activate this option to display assessors labels (full dendrogram) or classes (truncated dendrogram) on the dendrogram.
- **Colors**: Activate this option to use colors to represent the different groups on the full dendrogram.

Results

The **Products/Assessors table** is displayed, with the following elements:

- The first row displays the descriptor labels, which correspond to the same labels entered in the **Just about right data** field followed by the term "notenough", "JAR", or "toomuch".
- The first column contains the **Products** labels.
- Finally, the body of the table contains the fuzzy values (see section description).

Eigenvalues of CA: The eigenvalues of CA and corresponding chart (*scree plot*) are displayed.

Product coordinates: The coordinates of the products of the consensus in the factors space are displayed, with the corresponding charts (depending on the number of chosen factors).

Attribute coordinates: The coordinates of the attributes of the consensus in the factors space are displayed, with the corresponding charts (depending on the number of chosen factors).

Similarity matrix (S): The matrix of similarity index between all assessors is displayed. The similarity coefficient used is the Salton cosine which is included between 0 and 1. The closer it

is to 1, the stronger the similarity. This matrix is used by CATATIS to calculate the weights of the assessors.

Weights: The weights calculated by CATATIS are displayed, with the associated bar chart. The greater the weight, the more the assessor contributed to the consensus. Knowing that CATATIS gives more weight to the closest assessor from a global point of view, a much lower weight than the others will mean that the assessor is atypical.

Consensus configuration: The consensus configuration is displayed. It corresponds to the weighted average of the initial data.

Homogeneity: The homogeneity of the assessors is displayed. It is a value between $1/m$ (m being the number of assessors) and 1, which increases with the homogeneity of the assessors.

Similarity assessors/consensus: The similarity indices between the assessors and the consensus are displayed, with the associated bar chart. Like the weights of CATATIS, these coefficients make it possible to detect atypical assessors. The advantage of these coefficients is that they are between 0 and 1, so they are easier to interpret than the weights.

Global error: The global error of the CATATIS criterion is displayed. It corresponds to the sum of all residuals (which can be presented by assessor or product).

CLUSCATA-specific results:

Node statistics: This table shows the data for the successive nodes in the dendrogram. The first node has an index which is the number of assessors increased by 1. Thus, it is easy to see at any time if an assessor or group of assessors is clustered with another group of assessors in the dendrogram.

Levels bar chart: This table displays the statistics for dendrogram nodes, which correspond to the increase in the CLUSCATA minimization criterion (equivalent to the increase in within-class variance) when merging two classes.

Dendrograms: The full dendrogram displays the progressive clustering of assessors. If truncation has been requested, a broken line marks the level the truncation has been carried out. The truncated dendrogram shows the classes after truncation.

Example

A tutorial on how to analyze JAR data is available on the XLSTAT Help Center:

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

References

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.

RATA data analysis

Use RATA data analysis to analyze Rate-All-That-Apply (RATA) data. This method allows:

- To study and visualize the links between the products and attributes.
- To study the agreements between the assessors.
- to construct homogeneous groups of subjects.

In this section:

Description

The RATA test is a method used in sensory analysis to collect and analyse data on consumer perception of products.

RATA is useful for assessing sensory characteristics of products, informing product development and quality control in industries such as food and beverage, cosmetics, and consumer goods.

Participants evaluate product attributes, which may include taste, aroma, texture or appearance, using a numerical scale or rating system. Ratings are analysed to identify trends using multivariate statistical analysis.

In RATA data analysis, the fact that the subjects have had several sessions is allowed. In this case, the user chooses the option they prefer from the average per subject (if for a given product and a given attribute, he has scored 2 once and 0 the other time, the average of 1 will be taken into account) or to give dominance to the value 0 (if for a given product and a given attribute, he has scored 2 once and 0 the other time, the value 0 will be used, which implies that a subject indicating at least once that an attribute is not present definitively considers that this attribute is not present).

In addition, panel consistency tests overall and by attribute are proposed to determine if certain attributes are not included or give rise to excessively divergent responses. Tests on the weights to determine if some subjects have a non-significant weight can also be computed. This last option is particularly useful when dealing with experts.

RATA data analysis is a method that can be broken down into 3 main parts: * Performing an ANOVA for each of the attributes to check whether the effect of the product is significant;

- Using CATATIS to clarify the links between attributes and products.
- Using CLUSCATA to automatically construct groups of subjects with different points of view.

Uses of CATATIS

There are several applications for CATATIS, including:

- Study and visualization of the products and the attributes on the main planes.
- Study of the links between the assessors, especially to find the most atypical ones.

Principle of CATATIS

The goal of CATATIS is to form a consensus configuration that reflects at best the different assessors. This consensus can then be projected on different axes by a Correspondence Analysis or Principle Component Analysis (PCA). If the information associated with 2 or 3 first axes represents a sufficient percentage of the total variability of the consensus, the products and attributes will be able to be represented on a 2- 3-dimensional chart, thus making interpretation much easier.

Principle of CLUSCATA

The objective of CLUSCATA is to constitute classes of assessors as homogeneous as possible, each class of assessors being represented by a latent table (called consensus) determined by CATATIS. It is therefore natural that each class is finally analyzed by CATATIS, in order to determine the differences between the constituted classes. CLUSCATA consists of a hierarchical algorithm that can be "consolidated" by a partitioning algorithm (*i.e.* the partitioning algorithm is initialized by cutting the dendrogram). An interesting option is the creation of a "K+1" class (corresponding to an additional class) in order to set aside assessors who do not conform to any class. An assessor will be placed in this class if the similarities (Ochiai coefficients) between the consensus of each class and this assessor are all considered weak.

It should be noted that it is an adaptation of CLUSCATA made by our teams to non-binary data that is used.

Interpreting the results

The representation of the products and attributes in the space of k factors allows to visually interpret the proximities between the products and attributes, by means of precautions.

We can consider that the projection of a product or an attribute on a plan is reliable if it is far from the center of the graph.

Number of factors

Two methods are commonly used to determine how many factors must be retained for the interpretation of the results:

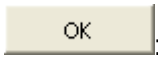
- Watch the decreasing curve of eigenvalues. The number of factors to be kept corresponds to the first turning point found on the curve.
- We can also use the cumulative variability percentage represented by the factor axes and decide to use only a certain percentage.

Graphic representations

These representations are only reliable if the sum of the variability percentages associated with the axes of the representation space are sufficiently high. If this percentage is high (for example 80%), the representation can be considered as reliable. If the percentage is low, it is recommended to produce representations on several axis pairs in order to validate the interpretation made on the two first factor axes.

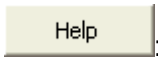
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.

: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.


: Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

RATA data: Select the data that correspond to the different assessors. If a column header has been selected, check that the "Attribute labels" option has been activated.

Product labels: Select the products corresponding to the RATA data. If a column header has been selected, check that the "Attribute labels" option has been activated.

Assessor labels: Select the assessors corresponding to the RATA data. If a column header has been selected, check that the "Attribute labels" option has been activated.

Sessions: Check this option if you want to use sessions in the analysis. If a column header has been selected, check that the "Attribute labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Attribute labels: Activate this option if the first row of the selected data (RATA data, Product labels, Assessor labels, Sessions) contains a header.

Analysis: Choose between one of the following two analyses:

- **CATATIS**: Choose this option to study and visualize links between products and attributes, and to study agreements between subjects using the CATATIS method.
- **CLUSCATA**: Choose this option to create homogeneous classes of subjects based on their perceptions of products, using the CLUSCATA method.

Options tab:

Filter factors: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %**: Enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Set the number of factors to take into account.

Confidence interval (%): Enter the confidence interval for the tests.

Number of permutations: Enter the number of permutations to be performed for the tests.

Session preprocessing: Choose the method to be used for session preprocessing.

Average: Activate this option if you want to use the average method.

Dominance of value 0: Activate this option if you want the value 0 to be considered as dominant, in the sense that if the judge/product/attribute triplet has the value 0 at least once for a session, then this value 0 will be retained for the rest of the calculations for this triplet.

CLUSCATA-specific options:

Truncation: Activate this option is you want XLSTAT to **automatically** define the truncation level, and therefore the number of classes to retain, or if you want to define the **number of classes** to create, or the **level** at which the dendrogram is to be truncated.

Consolidation: Activate this option to perform a consolidation of the classes obtained from the dendrogram.

Class K+1: Activate this option to add an additional class that will contain the assessors that do not fit any pattern of classes.

Rho parameter: Choose how you want to set the rho parameter: **automatically** or **user-defined**. This parameter represents the minimum agreement to be considered as sufficient in agreement to be kept in a class. The higher this parameter is set, the stronger the agreement required with the class and the more likely you are to place assessors in the K+1 class.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Replace missing data by 0: Activate this option to replace missing data by 0.

Outputs tab:

Similarity matrix (S): Activate this option to display the similarity matrix (Salton's cosine).

Weights: Activate this option to display the weights created and used by CATATIS.

Consensus configuration: Activate this option to display the consensus configuration created by CATATIS.

Similarity assessors/consensus: Activate this option to display the similarity coefficient between each assessor and the consensus.

Homogeneity: Activate this option to display homogeneity of the assessors.

Consistency tests: Activate this option to test if the panel is consistent globally and by attribute.

Global error: Activate this option to display the error of the CATATIS criterion.

Dimension reduction: Choose the method (CA or PCA) to be used to project the consensus configuration.

Eigenvalues: Activate this option to display the table and chart (scree plot) of eigenvalues of the CA or PCA on the consensus.

Coordinates: Activate this option to display the coordinates of the consensus in the factors space.

Zero Frequency: Activate this option to display the percentage of zeros present in the input data.

Pairwise Comparisons: Activate this option to display the results of Duncan's multiple comparison tests to facilitate detailed post-hoc analysis and comparison between groups.

CATATIS sub-tab:

Scaling factors: Activate this option to display the scaling factors.

Weight tests: Activate this option to test if the weights of the subjects are significant.

Residual per assessor: Activate this option to display the error of the CATATIS criterion for each assessor.

Residual per product: Activate this option to display the error of the CATATIS criterion for each product.

CLUSCATA sub-tab:

Node statistics: Activate this option to display the statistics for dendrogram nodes.

Class composition: Activate this option to display the composition of each class.

Charts tab:

Box plots: activate this option to display box plots showing basic statistics by attribute and by product.

Bar Chart: Enable this option to display the distribution of data by attribute.

Eigenvalues: Activate this option to display the *scree plot* of the CA or PCA eigenvalues.

Biplot: Activate this option to display the plot of the consensus coordinates in the factors space.

Display charts on two axes: Activate this option so that XLSTAT does not prompt you to select the axes, and automatically displays the biplot on the first two axes.

Scaling factors: Activate this option to display the bar chart of the scaling factors.

Weights: Activate this option to display the bar chart of the weights created and used by CATATIS.

Similarity assessors/consensus: Activate this option to display the bar chart of the similarity index between each assessor and the consensus.

Correlations charts: Activate this option to display charts showing the correlations between the components and initial variables. This chart is named correlation circle.

Residual per assessor: Activate this option to display the bar chart of the error of the CATATIS criterion for each assessor.

Residual per product: Activate this option to display the bar chart of the error of the CATATIS criterion for each product.

CLUSCATA sub-tab:

Levels bar chart: Activate this option to display the diagram of levels showing the impact of successive clusterings on the within-class variance.

Dendrogram: Activate this option to display the dendrogram.

- **Horizontal**: Choose this option to display a horizontal dendrogram.
- **Vertical**: Choose this option to display a vertical dendrogram.
- **Truncated**: Activate this option to display the truncated dendrogram (the dendrogram starts at the level of the truncation).
- **Labels**: Activate this option to display assessors labels (full dendrogram) or classes (truncated dendrogram) on the dendrogram.
- **Colors**: Activate this option to use colors to represent the different groups on the full dendrogram.

Results

Zero Frequency: The table displaying the percentage of zeros present in the input data.

ANOVA summaries: The ANOVA summary for each attribute is displayed.

Assessors' repeatability: The coefficient of similarity (Salton's Cosine) between the results of different sessions is displayed. This coefficient takes values between 0 and 1 and increases with the similarity between sessions.

Eigenvalues: The eigenvalues of CA or PCA and corresponding chart (*scree plot*) are displayed.

Product coordinates: The coordinates of the products of the consensus in the factors space are displayed, with the corresponding charts (depending on the number of chosen factors).

Attribute coordinates: The coordinates of the attributes of the consensus in the factors space are displayed, with the corresponding charts (depending on the number of chosen factors).

Similarity matrix (S): The matrix of similarity index between all assessors is displayed. The similarity index is included between 0 and 1. The closer it is to 1, the stronger the similarity. This matrix is used by CATATIS to calculate the weights of the assessors.

Scaling factors: The scaling factors are displayed with the associated bar chart. The larger a scale factor of an assessor, the smaller the original rate scale of this assessor.

Weights: The weights calculated by CATATIS are displayed, with the associated bar chart. The greater the weight, the more the assessor contributed to the consensus. Knowing that CATATIS gives more weight to the closest assessor from a global point of view, a much lower weight than the others will mean that the assessor is atypical.

Weight tests: The results of the weight tests are displayed. If a subject has a non-significant weight, then his point of view is very different from the global point of view, and his results can be questioned if he is an expert.

Consensus configuration: The consensus configuration is displayed. It corresponds to the weighted average of the initial data.

Homogeneity: The homogeneity of the assessors is displayed. It is a value between $1/m$ (m being the number of assessors) and 1, which increases with the homogeneity of the assessors.

Consistency tests: The results of the consistency tests are displayed globally and by attribute. If the panel is globally inconsistent, the data can unfortunately be discarded. If it is inconsistent for one or more attributes, then those attributes are subject to so much disagreement that they have surely been misunderstood.

Distance between the median of permutations and homogeneity: This distance shows how strong the homogeneity of subjects is compared to random responses.

Similarity assessors/consensus: The similarity indices between the assessors and the consensus are displayed, with the associated bar chart. Like the weights of CATATIS, these coefficients make it possible to detect atypical assessors. The advantage of these coefficients is that they are between 0 and 1, so they are easier to interpret than the weights.

Global error: The global error of the CATATIS criterion is displayed. It corresponds to the sum of all residuals (which can be presented by assessor or product).

Residual per assessor: This table and the corresponding bar chart make it possible to visualize the distribution of the residuals by assessor. It is thus possible to identify for which assessors CATATIS has been less efficient, or in other words, which assessors stand out the most from the consensus.

Residual per product: This table and the corresponding bar chart make it possible to visualize the distribution of the residuals by product. It is thus possible to identify for which products CATATIS has been less efficient, or in other words, which products stand out the most from the consensus.

CLUSCATA-specific results:

Node statistics: This table shows the data for the successive nodes in the dendrogram. The first node has an index which is the number of assessors increased by 1. Thus, it is easy to see at any time if an assessor or group of assessors is clustered with another group of assessors in the dendrogram.

Levels bar chart: This table displays the statistics for dendrogram nodes, which correspond to the increase in the CLUSCATA minimization criterion (equivalent to the increase in within-class variance) when merging two classes.

Dendrograms: The full dendrogram displays the progressive clustering of assessors. If truncation has been requested, a broken line marks the level the truncation has been carried out. The truncated dendrogram shows the classes after truncation.

If **multiple comparison** tests have been requested, the results corresponding are then displayed.

If several dependent variables have been selected and the option to multiple comparisons has been activated, a table showing the means of each category of each factor and for all Y is

displayed.

Example

A tutorial on how to use the RATA data analysis feature is available on the XLSTAT Help Center:

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

References

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. (2020). Classification de tableaux de données, applications en analyse sensorielle (Doctoral dissertation, Nantes, Ecole nationale vétérinaire).

Llobell, F., Bonnet, L., & Giacalone, D. (2024). Assessment of panel performance in CATA and RATA experiment. Journal of Sensory Studies, 39(4), e12941.

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., Jaeger, S.R. (September 11, 2024). Consumer segmentation based on sensory product characterisations elicited by RATA questions? Eurosense conference, Dublin, Ireland.

Flash Profiling

Use the Flash Profiling to analyze the data from this test, in the form of several tables of product/attribute data. In this way, you'll be able to obtain results that can be used not only to characterize your products, but also to define a sensory map.

In this section:

Description

The Flash Profiling is a sensory analysis method proposed by Dairou & Sieffermann in 2002, which combines a free profile technique (free choice of attributes) with a ranking of products for each attribute.

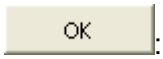
Faced with a set of products, untrained assessors choose their own terms to describe and distinguish the main sensory differences between the different products. These same assessors then rank the products for each of the attributes they have chosen.

The advantage of the Flash Profiling is that it does not require assessors to be trained, as they are only asked to rank the products on an ordinal scale for each of the attributes, which is different from a free profile data analysis where assessors have to rate the intensity of the different attributes.

In this analysis, the data first undergoes an ANOVA for each attribute on ranks, which highlights perceived differences between products and their discrimination. A Multiple Factor Analysis (MFA) is then performed on the data to determine the consensus, the sensory map and the assessors' agreement with the consensus.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to products and columns to attributes. If the arrow points to the right, XLSTAT considers that rows correspond to attributes and columns to products.

General tab:

Products/attributes table: Select the data that correspond to N products described by P attributes for K assessors. If column headers have been selected, check that the "Variable labels" option has been activated. If the data is **not suitable** for this type of table, it can be **transformed** into a horizontal table, as described above, using the **Create a Products Table.** feature.

Number of assessors: Enter the number K of assessors in which the selected data are subdivided.

Assessor labels: Activate this option if you want to use labels for the K assessors. If this option is not activated, the name of the tables are automatically generated (Assessor1, Assessor2, etc.). If column headers have been selected, check that the "Variable labels" option has been activated.

Number of attributes per assessor:

- **Equal**: Choose this option if the number of attributes is identical for all the assessors. In that case XLSTAT determines automatically the number of attributes for each assessor.
- **User defined**: Choose this option to select a column that contains the number of assessors for each assessor. If the "Variable labels" option has been activated, the first row must correspond to a header.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (Products/attributes table, products labels) includes a header.

Products labels: Activate this option if products labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the products labels are automatically generated by XLSTAT (Product1, Product2 …).

Options tab:

Filter factors: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %**: Activate this option then enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Activate this option to set the number of factors to take into account.

Display charts on two axes: Activate this option if you want the numerous graphical representations displayed are only displayed on the first two axes, without your being prompted after each analysis.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data has been detected.

Remove observations: Activate this option to ignore the products that contain missing data.

Adapted strategies: Activate this option to choose strategies that are adapted.

- **Mean**: Activate this option to estimate the missing data of a product by the mean of the corresponding attribute.
- **Nearest neighbor**: Activate this option to estimate the missing data of a product by searching for the nearest neighbor of the product.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics for all the selected attributes.

Eigenvalues: Activate this option to display the table and chart (scree plot) of eigenvalues.

Contributions: Activate this option to display the contribution tables.

Squared cosines: Activate this option to display the tables of squared cosines.

Assessors:

- **Coordinates**: Activate this option to display the coordinates of the assessors in the MFA space. Note: the contributions and the squared cosines are also displayed if the corresponding options are checked in the Outputs/General tab.
- **Lg coefficients**: Activate this option to display the Lg coefficients.
- **RV coefficients**: Activate this option to display the RV coefficients.

Attributes:

- **Factor loadings**: Activate this option to display the factor loadings in the MFA space.
- **Attributes/factors correlations**: Activate this option to display the correlations between factors and attributes in the MFA space.

Products:

- **Factor scores**: Activate this option to display the factor scores in the MFA space.
- **Coordinates of the projected points**: Activate this option to display the coordinates of the projected points in the MFA space. The projected points correspond to the projections of the products in spaces reduced to the number of dimensions of each assessor.

Charts tab:

Assessors charts: Activate this option to display the charts that allow to visualize the assessors in the MFA space.

Correlation charts: Activate this option to display the charts involvingcorrelations between the components and the attributes used in the MFA.

Products charts: Activate this option to display the chart of the products in the MFA space.

Charts of the projected points: Activate this option to display the chart that shows at the same time the products in the MFA space, and the products projected in the sub-space of each assessor.

- **Products labels**: Activate this option to display the products labels on the charts.
- **Projected points labels**: Activate this option to display the labels of the projected points.

Color by group: Activate this option, if you want to color product points according to levels of a qualitative variable. Then select a vertical vector that must have as many rows as there are active products. If headers were selected for the main table, ensure that a label is also present for the variable in this selection.

- **Confidence ellipses**: Activate this option if you want to display confidence ellipses around group of products corresponding to the levels of the group variable selected to color products. You also have to select the confidence interval for the ellipses.

Options for attributes:

Filter: Activate this option to modulate the number of attributes displayed:

- **Random**: The attributes to display are randomly selected. The "Number of attributes" N to display must then be specified.
- **N first columns**: The first N attributes are displayed on the chart. The "Number of attributes" N to display must then be specified.
- **N last columns**: The last N attributes are displayed on the chart. The "Number of attributes" N to display must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the attributes to display.
- **Sum(Cos2)>**: Only the attributes for which the sum of squared cosines (communalities) are bigger than a value to enter are displayed on the plots.

Options for products:

Filter: Activate this option to modulate the number of product displayed:

- **Random**: The products to display are randomly selected. The "Number of products" N to display must then be specified.
- **N first rows**: The first N products are displayed on the chart. The "Number of products" N to display must then be specified.
- **N last rows**: The last N products are displayed on the chart. The "Number of products" N to display must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the products to display.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the attributes selected. This includes the number of products, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased).

The **ANOVA Summaries** table shows the different p-values of the various ANOVAs by attribute, and thus highlights the differences perceived by the assessors between the different products and their discrimination. The attribute must be present at least twice in the dataset for the ANOVA to be performed.

Afterwards, the results of the MFA are displayed. First, eigenvalues, attribute results and product results are displayed.

At the end of the products coordinates table, the following button is displayed:  . Click on this button to automatically open the pre-filled dialog box of HAC (Hierarchical Ascending

Classification) and perform a classification of the products on the factorial coordinates.

The **coordinates of the assessors** are then displayed and used to create the plots of the assessors. The latter allow to visualize the distance between the assessors. Then contributions and squared cosines for assessors are displayed.

Lg coefficients: The Lg coefficients of relationship between the assessors allow to measure to what extent the assessors are related two by two. The more attributes of a first assessor are related to the attributes of the second assessor, the higher the Lg coefficient.

RV coefficients: The RV coefficients of relationship between the assessors are another measure derived from the Lg coefficients. The value of the RV coefficients varies between 0 and 1.

The **coordinates of the projected points** in the space resulting from the MFA are displayed. The projected points correspond to projections of the products in the spaces reduced to the dimensions of each assessor. The representation of the projected points superimposed with those of the complete products makes it possible to visualize at the same time the diversity of the information brought by the various assessors for a given product, and to visualize the relative distances from two products according to the various assessors.

*Remark about the **Axes homogeneity index***: This index developed by our team is very useful to determine if the contributions of the products are homogeneous for the different axes. It is constructed as the proportion of products with an absolute contribution $> 1/n$. An index above 0.4 indicates a very good homogeneity with well represented products. On the other hand, an index lower than 0.1 should be a warning to the user who should check if there are no outliers in the attributes constructing the axis that would distort its interpretation (the outliers would then be the products that stand out from the others on the axis in question).

The **ranks table** shows the rankings assigned to each product for the various attributes. The product with the lowest intensity on an attribute will have rank 1 for that attribute. Note that in the event of a tie between several products for a given attribute, the rank assigned to these products corresponds to the average of the ranks of the tied products.

Example

An example of a Flash Profiling analysis is available in the XLSTAT Help Center at the following address:

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

References

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).

Dairou V., Sieffermann J-M. (2002). A comparison of 14 jams characterized by conventional profile and a quick original method, the Flash Profile. *Journal of Food Science*, **67** (2), 826–834.

Delarue J., Sieffermann J-M. (2004). Sensory mapping using Flash profile. Comparison with a conventional descriptive method for the evaluation of the flavour of fruit dairy products. *Food*

Quality and Preference, **15** (4), 383-392.

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.

Marketing tools

Sample size

Use this tool to calculate the number of respondents needed to obtain statistically strong results for a population or to obtain the margin of error for your sample.

In this section:

Description

When talking about a statistical study or a market study involving research on groups of populations, it becomes important to ask the question of the sample size. In order to obtain good results, it is necessary to have a sample that is as representative as possible of the population. XLSTAT makes it possible, with this tool, to calculate the right number of people to be interviewed in order to obtain neither too large (which would make the study more complex and expensive), nor too small (which would generate erroneous results). If you already know the size of your sample, XLSTAT also allows you to check its margin of error.

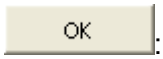
The following formula is used to calculate the results:

$$\text{Sample size} = \frac{Z^2 * p * (1-p)}{(\text{margin of error})^2}$$

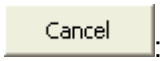
where Z is the score of the normal distribution defined from confidence interval entered in the interface, p is the proportion of the population presenting the characteristic studied (for the sake of simplicity, it is automatically chosen at 0.5 in XLSTAT), and the margin of error is the difference you accept between the sample mean and the population mean.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. Below are descriptions of the various elements of the dialog box.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help options.


: Click this button to reload the default options.

General tab:

Goal*: Choose between the calculation of the sample size and the calculation of the margin of error (depending on this choice the following fields will be different).

Population Size: Enter the size of the study population.

Margin of error (in case of sample size calculation): Enter the margin of error that you accept for your study (in %).

Number of respondents (in the case of calculating the margin of error): Enter the number of people who responded to your survey.

Confidence interval: Enter the size of the desired confidence interval (in %).

Estimated response rate: Select this option if you want to calculate the number of invitations required to reach the correct sample size. Then enter the estimated response rate (in %).

Calculate: Click this button to display the result in the *Results* area of the dialog box.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Results: Results are displayed in this area.

Clear: Click this button to clear the results saved in the *Results* area of the dialog box.

Results

The results displayed by XLSTAT are a summary table with the population size, the margin of error, the confidence interval, the sample size and the number of invitations required.

Waffle Chart: Allows to represent the percentage of the sample size compared to the size of the population.

Pie Chart: Used to plot the sample size against the number of invitations required.

Example

An example of using the sample size calculator is available on XLSTAT Help Center:

https://www.xlstat.com/demo/samplesize_en

Price Sensitivity Meter (Van Westendorp)

Use this tool to identify the acceptable price range or the optimal prices for trial or revenue, based on survey data.

In this section:

Description

This version of price sensitivity analysis (and its impact on sales volume and revenue) is due to Van Westendorp and was presented at the ESOMAR congress in 1976. It was later enriched by Newton *et al.* in 1993, to respond to mainstream criticism that this previous method did not take purchase intention into account.

This method consists of conducting a survey on a group of panelists of consumers and asking them how they perceive the price of a specific product. + TCH: *too cheap* (the price point is perceived too low for this specific product) + CH: *cheap* (the lower price range at which one might buy the product) + EX: *expensive* (the upper price range at which one might buy the product) + TEX: *too expensive* (above the price one would be willing to pay for this specific product)

For this type of price sensitivity analysis, the usual result is a graph presenting a series of curves whose intersections determine critical prices: from the survey data, six cumulative distribution curves (or their opposites) are computed, first for the four types of prices collected (*too cheap*, *cheap*, *expensive*, *too expensive*), then, by deduction, we calculate the distribution for *not cheap* and *not expensive*.

For *too cheap* and *cheap*, we take the opposite of the distribution curve. For each of the prices indicated by the panel participants, we calculate what proportion of respondents s indicated a higher price. These curves therefore decrease from 1 to 0 when the price increases. For *expensive* and *too expensive*, we take the cumulative distribution curve. Therefore, for each of the prices indicated by the respondents, we calculate what proportion of consumers indicated a lower price. These curves therefore increase from 0 to 1 when the price increases.

The intersection between the cheap and expensive curves is the price for which the same number of panelists consider the product to be *expensive* or *cheap*. Even if it is not necessarily strong, there is a disagreement between these two groups of same size of respondents on this price which has been named **Indifference Price (IDP)**. According to Van Westendorp, this price corresponds to the reality of the market. The IDP can be interpreted as the median market price for this type of product or as the price offered by a market leader.This is the right price point for a majority of respondents , a small proportion finding it cheap (probably not expensive enough

for the company marketing this product) or *expensive* (potentially at risk for of a concurrent). Of course, an even smaller respondents will find it *too cheap* (suspicious quality) or *too expensive* (inaccessible).

The intersection between the *too cheap* and *too expensive* curves is the price at which many respondents consider the product to be *too cheap* or *too expensive*. This price which corresponds to radically opposed opinions, concerns a tiny group of panelists. This price is called the **Optimal Pricing Point (OPP)**, price point at which the purchase intent will not be impacted negatively, meaning that outside of any other external factor, few people would be discouraged by this price.

The acceptable price range is given by, for the lower bound, the intersection between the *too cheap* and *not cheap* curves, and for the upper bound, by the intersection of the *too expensive* and *not cheap* curves. Between these two marginal prices, the authors estimate that sales volumes are high.

Purchase intent

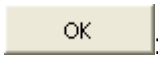
XLSTAT allows to take into account the contribution of Newton *et al.* (1993) who proposed to take into account the purchase intent within the framework of the survey, by asking what is the purchase intent score, for the *cheap* and *expensive* prices. These scores can be transformed into probabilities, either automatically or through a conversion table. Once probabilities are available, we can identify which price is likely to generate a maximum volume of sales and which price is likely to generate a maximum revenue.

Data format

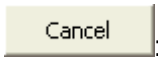
XLSTAT allows producing partial results if only the *cheap* and *expensive* prices are available. It is however recommended that you survet the four prices (TCH/CH/EX/TEXT) to take full advantage of the method.

Dialog box

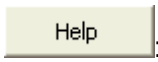
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Price data (TCH/CH/EX/TEX): Select the data that correspond to prices. You must either select two columns (CH/EX) or four columns (TCH/CH/EX/TEX). If a column header has been selected, check that the "Variable labels" option has been activated.

Check consistency: Activate this option if you want XLSTAT to check that prices are in ascending order for each individual. Otherwise, the individual is not taken into account for the analysis.

Groups: Activate this option if you want the perform the analyses per group, and then select the data that indicate to which group each individual belongs. If a column header has been selected, check that the "Variable labels" option has been activated.

Purchase intent data (CH/EX): Activate this option if you surveyed the consumers for their purchase intent at the *cheap* and *expensive* prices. These data can be scores (for example on a 1-5 scale) or probabilities If a column header has been selected, check that the "Variable labels" option has been activated. If this option is activated and the purchase intent data are not probabilities and the conversion table is not provided, XLSTAT automatically performs the conversion into probabilities.

Conversion table: Activate this option if you want XLSTAT to convert the purchase intent scores into probabilities using your input.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Fit normal distributions: Activate this option if you want to fit normal distributions to each price sample and to use cumulative normal distribution to compute the different prices and price

ranges.

Outputs tab:

Individual results: Activate this option to display the purchase probabilities for each individual at the optimal trials and optimal revenue prices.

Results

Summary statistics: This table displays descriptive statistics for all the prices selected and for their range. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed.

Box plots: Box plots are then displayed so that you can visualise in one look the distribution of the different prices. If you the individuals belong to different groups defined in the dialog box, for each price (TCH/CH/EX/TEX), a separate chart allows to compare the groups.

Statistics: This table displays the **Indifference Price (IDP)** and the **Optimal Pricing Point (OPP)**. The price and the proportion of consumers on the corresponding cumulative curves at this point, are displayed.

The **acceptable price range** is then displayed.

The **Price Sensitivity Meter** chart is the main result of the method as it displays the different cumulative price curves. On the same chart, XLSTAT displays the IDP, the OPP and the acceptable price range.

If the purchase intent data has been entered, the next table shows the optimal trial price and optimal revenue price, as well as the respective average probability of purchase, expected volume on the population of the study and the expected revenue.

Volume and revenue curves based on prices are also displayed. The last table shows, if the option is activated in the dialog box, the probability of purchase for the optimal trials prices and turnover.

Example

An example showing how to use the *Price Sensitivity Meter* is available on the XLSTAT Help Center at

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

References

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.

Price elasticity of demand

Use this tool to compute the price elasticity of demand using as input data the quantities sold at different prices and identify the price at which the maximum revenue is generated.

In this section:

Description

The analysis of **price elasticity of demand** (PED or E_d) is essential in marketing because it is an approach that allows setting the price of a product, knowing that the goal, at least in the mid term, is to maximize the revenue generated by the product.

The elasticity which concept is due to Alfred Marshall (1920) is defined as the relative variation of demand (or quantity sold) Q when price P changes. The definition writes:

$$E_d = \frac{dQ/Q}{dP/P} \tag{1}$$

In general, an increase in prices is accompanied by a decrease in the quantities sold (although many examples to the contrary exist). Elasticity is therefore often a negative quantity, although some authors take the opposite.

As dQ/dP mathematically corresponds to an infinitesimal variation at a given point, in practice, we compute the *point elasticity* or the *arc elasticity*.

Point elasticity is defined as ratio of the relative variation of the quantity sold when prices are changed from $P1$ à $P2$, and the relative variation of prices.

$$E_d = \frac{(Q2 - Q1)/Q1}{(P2 - P1)/P1} \tag{2}$$

Notes: + We say that demand is price elastic if the demand decreases strongly when price increases from $P1$ to $P2$ ($P1Q1 > P2Q2$). The term elastic is used because quantities "react" strongly to the change as if they were linked to prices with an elastic. ($E_d < -1$). It is perfectly elastic if $E_d = -\infty$. + Unit elastic with $E_d = -1$ if demand evolves like prices. When elasticity is -1, there is no impact on the revenue: the drop in the quantity sold is exactly compensated by the increase in prices. Typically, a 1% increase in prices, leads to a 1% decrease of quantities, maintaining the revenue at the same level. + We say that demand is

inelastic (or rigid) if the increase of price has a relatively lower impact on demand ($-1 < E_d \leq 0$). Demand is perfectly inelastic if $E_d = 0$. + $E_d > 0$ happens for some specific products (Giffen goods which are essential goods, or Veben goods such as luxury goods). The increase of prices have a significant positive effect on quantities. + One cannot deduce the evolution of revenue from the price elasticity.

A problem with *point elasticity* is that it is not the same if the reference price $P1$ is the lower or higher price. To avoid that problem, the *arc elasticity* is often prefered.

Arc elasticity is the ratio of, the relative variation of the quantity sold $Q2$, when prices increase from $P1$ to $P2$, and the relative variation of prices.

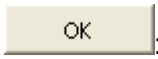
$$E_d = \frac{(Q2 - Q1)/((Q1 + Q2)/2)}{(P2 - P1)/((P1 + P2)/2)} = \frac{(Q2 - Q1)/(Q1 + Q2)}{(P2 - P1)/(P1 + P2)} \quad (3)$$

XLSTAT allows to compute both types of elasticities.

It is common to represent the curves connecting quantity and price by putting the price in the ordinate and the quantity in the abscissa. This is counterintuitive in our case, since we are here in a perspective where the quantities sold depend on prices. The use of quantities in ordinates can be explained by the underlying economic theory of the study of supply and demand and the study of market equilibria, where the price is determined according to available supply and demand. XLSTAT displays the graphics following both alternatives.

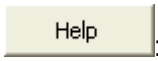
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.

: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Prices: Select the data that correspond to prices. If a column header has been selected, check that the "Variable labels" option has been activated.

Demand: Select the data that correspond to the demand (quantities sold) corresponding to each price. If a column header has been selected, check that the "Variable labels" option has been activated.

Groups: Activate this option if you want the perform the analyses per group, and then select the data that indicate to which group each individual belongs. If a column header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections includes a header.

Point elasticity: Activate this option to compute and display *point elaticities*.

Arc elasticity: Activate this option to compute and display *arc elaticities*.

Results

Summary statistics: This table displays descriptive statistics for prices and demand. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed.

XLSTAT displays in a first table the price, the demand, the revenue and the point elasticity if it was requested. If the arc elasticity has been requested, a second table displays for each median point calculated the price, the demand, the revenue and the arc electricity.

These various elements are then crossed in a series of graphs.

Example

An example showing how to compute the *Price elasticity of demand* is available on the XLSTAT Help Center at

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

References

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)

Customer Lifetime Value (CLV) will help you to estimate the cash flows you will get from customers and estimate how long you can keep them after the acquisition. Using CLV, you will be able to streamline the marketing or advertising operation you could engage in order to increase your customers retention rate.

In this section:

Description

How much profit do you expect to make from your customers during their lifetime? And how would increasing retention rates affect future profit? These are among the most common questions CLV could answer.

Customer Lifetimes Value (CLV): Customer Lifetime Value (CLV) can be defined as the present value of the future cash flows attributed to the relationship between a customer and the company.

Several models for calculating CLV exist which vary according to multiple factors, specific to each organization. Here,the implemented model is the Simple Retention Model (SRM), suitable for customers in a contractual situation. The Simple Retention Model (SRM) estimates CLV assuming the following conditions:

- The percentage of retained customers each period (retention rate r) is constant over time and across customers.
- The period cash flow m is unaffected by cancellation time.
- The event that a customer cancels the subscription at a time period t is independent of the event that the customer cancels during any other period.

Probabilistic model for CLV:

- **Cancellation/churn time**: Assume that all customers, being part of a certain segment, are retained each subscription period with probability r (retention rate) for all periods. Moreover, assume that a customer cancellation occurring at a certain period is independent of the cancellation during any other period. Let T be a random variable indicating the time of cancellation and t be a realization of T. Under these assumptions, T follows a geometric distribution. The probability for a geometric distribution is given by:

$$f(t) = P(T = t) = r^{t-1}(1 - r)$$

This formula gives the probability that a customer cancels at time t. It can be also interpreted as the probability that a customer survives $t - 1$ periods.

XLSTAT also calculates the quantiles of T. The α quantile of the random variable T, called P_α divides the distribution of the random variable T such that α percent of the distribution has $T \leq P_\alpha$ and $1 - \alpha$ percent of the distribution has $T \geq P_\alpha$. So we have $P(T \leq P_\alpha) = \alpha$ and $P(T \geq P_\alpha) = 1 - \alpha$. Under the assumptions of the SRM we have:

$$P_\alpha = \frac{\log(1 - \alpha)}{\log r}$$

- **CLV**: When a customer cancels during period t, there will be t cash flows if it occurs at the beginning of the period and $t - 1$ cash flows if it is at the end of the period. Let d be the discount rate. For a given cancellation time, we can compute the CLV using the following formulas:

$$CLV = \sum_{t=0}^{T-1} \frac{m}{(1+d)^t} = m \times \frac{(1+d)[1-(1+d)^{-T}]}{d} \quad \text{payment: start of period}$$

$$CLV = \sum_{t=1}^{T} \frac{m}{(1+d)^t} = m \times \frac{1-(1+d)^{-T}}{d} \quad \text{payment: end of period}$$

However, the cancellation time T is random, and CLV will thus follow a distribution. Customers with larger T have a larger CLV. Therefore we can summarize the distribution of the CLV with its mean or expectation.

$$E[CLV] = \frac{m(1+d)}{1+d-r} \quad \text{payment: period start}$$

$$E[CLV] = \frac{m \times r}{1+d-r} \quad \text{payment: period end}$$

- **Estimation of retention rates**: In the previous section, we assume that the retention rate r was known, but in practice it's not always the case. That's why XLSTAT proposes to estimate the later from the data. In any organization, some but not all customers will cancel. A customer who has not yet canceled is said to be censored. Thus, we can say that the organization will not have yet observed this customer cancellation time.

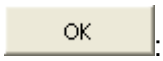
Let n_0 be the number of customers who have not yet canceled and n_1 the number of customers who have already canceled so that cancellation/defection time t has been observed.

If customer i has already canceled, let t_i be the observed time of defection, so that customer i has been active during $t_i - 1$ periods. Let C_i be the time of censoring for customers who are still active. For those customers we have $T_i > C_i$. The retention rate is estimated using the following formula:

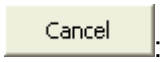
$$\hat{r} = 1 - \frac{n_1}{\sum_i t_i + \sum_i C_i}$$

Dialog box

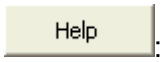
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

ARPA (Average Revenue Per Account): Select the data that correspond to revenues. You must select only one column. If a column header has been selected, check that the "Variable labels" option has been activated.

Time (acquisition/churn): Select in order the two columns: acquisition date and churn date. If a customer has not yet left, the second column of date should be an empty cell. If a column header has been selected, check that the "Variable labels" option has been activated.

Segments: Activate this option if you want to perform the analyses per segment, and then select the data that indicate to which segment each customer belongs. If a column header has been selected, check that the "Variable labels" option has been activated.

Subscription period: Select the subscription period.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections includes a header.

Account names: Activate this option if accounts labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Options tab:

Segments labels: Select the segments labels. This field is only available when the segments option is enabled and at least one of the following options is enabled: discount rate, fixed costs or user defined retention rate.

Discount rate: Activate this option if you want to take into account the discount rate applied to customers. This rate is considered fixed. If the segments option is enabled, select one value per segment and make sure that the lines are in the same order as segments labels.

Fixed costs: Activate this option if you want to deduct certain fixed operating costs from the revenues generated when calculating CLV. Select multiple columns if you have multiple costs to include. If the segments option is enabled, select one value per segment and make sure that the lines are in the same order as segments labels.

Retention rate:

- **Estimate**: Activate this option to estimate the retention rate from the input data. If the segments option is enabled, the retention rate will be estimated for each segment.
- **User defined**: Activate this option if you want to define the retention rate yourself. This rate is considered fixed. If the segments option is enabled, select one value per segment and make sure that the lines are in the same order as segments labels.

Payment:

- **Start of period**: Select this option if payments occur at the beginning of each period.
- **End of period**: Select this option if payments occur at the end of each period.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the revenue variable (ARPA).

CLV: Activate this option to display the average CLV. If the segment option is enabled, this value is displayed for each segment.

Estimated churn rate: Activate this option to display the churn and retention rate. If the segment option is enabled, these values are displayed for each segment.

Estimated defection time: Activate this option to display statistics on the time before defection in order to quickly view the dispersion of customers cancellation times.

Cashflow evolution: Activate this option to display the cash flow evolution. Each line in this table contains a cash flow forecast by period.

Individual results: Activate this option to display the CLV by customer.

CLV forecast: Activate this option if you want to make simulations on the CLV.

- **duration**: Choose the period for which you want to perform the simulation.

Sensitivity analysis: Activate this option to display he impact of an increase in retention rate on CLV.

Charts tab:

CLV per segment: Activate this option to display the CLV by segment.

Estimated defection time: Activate this option to display the chart summarizing the informations contained in the corresponding table described above.

Churn probabilities: Activate this option to display the chart showing the churn probabilities as a function of time.

Cashflow evolution: Activate this option to display the chart showing the cashflow evolution as a function of time.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics. The number of Observations, the minimum and maximum values, the quartiles, the mean, the standard deviation and the variance (unbiased) are displayed for the revenues variable (ARPA).

Box plots (or box-and-whisker plots) related to the revenues variable (ARPA) are also displayed (see description).

CLV: The average CLV is displayed. If the segment option is enabled, this value is displayed for each segment.

Estimated churn rate: In this table are displayed the churn and retention rate. If the segment option is enabled, these values are displayed for each segment.

Estimated defection time: Statistics on the time before defection are displayed in order to quickly visualize the dispersion of customers cancelation times. Thus, the 1st quartile, 3rd

quartile, median and mean of customers departure times are displayed. If the segment option is enabled, these values are displayed for each segment.

Cashflow evolution: This table shows the evolution of cash flow. Each line in the table contains a cash flow forecast by period. The first line corresponds to the period following the most recent period contained in the input data.

Individual results: The CLV is displayed for each customer.

CLV forecast: A simulation on the average CLV value of customers remaining in the database after the last recorded churn date is performed over the period chosen by the user.

Sensitivity analysis: The impact of an increase in retention rate on CLV is displayed. The variations considered are increments of 5% from the estimated/defined retention rate. Each line in the table corresponds to a simulated retention rate. The CLV and the average time before defection are displayed in columns.

The **CLV per segment** chart is only available when the segment option is active. This graph displays the CLV for each segment as a bar chart.

Estimated defection time: This chart summarizes in a bar chart the information contained in the corresponding table described above.

Churn probabilities: Activate this option to display the chart showing the churn probabilities as a function of time.

Cashflow evolution: Activate this option to display the chart showing the cashflow evolution as a function of time.

Example

An example showing how to use the *Customer Lifetime Value* is available on the XLSTAT Help Center at:

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

References

Phillip E. Pfeifer, Mark E. Haskins and Robert M. Conroy. Customer Lifetime Value, Customer Profitability, and the Treatment of Acquisition Spending. *Journal of Managerial Issues, Vol. 17, No. 1 (Spring 2005), pp. 11-25*.

Malthouse, Edward C. (2013). Segmentation and Lifetime value Models Using SAS®. *Cary, NC: SAS institute Inc*.

Customer Long-term Value (CLTV)

Based on your order history, customer long-term value (CLTV) will help you to estimate the cash flows you will get from customers and estimate how long you can keep them after the acquisition. It will also allow you to better understand your customers life cycle, identify periods of high churn risk, and gives you an estimation of the profits generated by your customers over an extended time period.

In this section:

Description

How much profit do you expect to make from your customers during their lifetime? And how would increasing retention rates affect future profit? These are among the most common questions CLV could answer.

What is the distribution of events (cancelation) during the lifetime of your customers? Or when customers are more likely to churn ? These are some of the questions that this module (CLTV) could help you answer.

Customer Long-term Value (CLTV):

Customer Lifetime Value (CLV) can be defined as the present value of the future cash flows attributed to the relationship between a customer and the company.

Here, the implemented model is the General Retention Model (GRM), this model as the simple retention model (have a look on the method description), is suitable for customers in a contractual situation.

The simple retention model (SRM) assumes that the retention rate is constant over time. But in many cases, retention rates are not always constant. For example, telephone companies or internet service provoiders commonly offer a few months of free acces or a few moths at very low price and then increase it. In such cases, the retention rate often begins high and then drop after the discount period.

The general retention model (GRM) extend the SRM by allowing retention rates r to vary over time and cash flows m to depend on the time of cancelation.

However, we still assume that the event that a customer cancels the subscription at a time period t is independent of the event that the customer cancels during any other period.

As opposed to the SRM, in which the CLV can be estimated in perpetuity due to the constant retention rate, here the CLV will be estimated from what has been observed in your order history. If the customers history extends over a 10-year period then the model will be able to produce an estimate of the CLV over 10 years based on your customers history.

Using general retention model (GRM) to compute CLTV :

To implement this model, we will use methods related to survival analysis, more specifically the life table analysis (see the section description of the method).

- **Retention function**: The survival function, which in this case correspond to the retention function, gives for each time period t the chance that the customer is retained for the $t-1$ periods.

Let R_t be the event that "the customer is retained in period t". Using independance, we have:

$$P(R_1 \bigcap R_2) = P(R_1) \times P(R_2) = r_1 \times r_2$$

By extending this reasoning to the first $t-1$ periods, we get the retention function above:

$$S(t) = P(T \geq t) = \prod_{i=1}^{t-1} r_i$$

Note that the churn function correspond to $1 - S(t)$.

- **Probability density function**: the probability density function gives the probability that a customer is retained during the first $t-1$ periods and cancels during period t. We have:

$$f(t) = P(t = T) = S(t)(1 - r_t) = S(t) - S(t+1)$$

- **Hazard rate**: let π_t be the hazard rate at time t. It's the conditionnal probability of canceling at time t given that the customer has not already canceled:

$$\pi_t = P(T = t | T \geq t) = \frac{P(T=t)}{S(t)} \simeq 1 - r_t$$

Note that hazard rates are not always probabilities, but here we consider discrete time intervals, so they are.

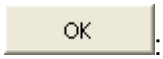
- **Compuation of CLV**: let m_t be the cash flow (eventuals discounts included) at time t. For a customer who cancels at time $T = t$, we have $CLV = \sum_{t=0}^{t-1} m_i$. Here we would like to find the expected value of $CLV(T)$.

$$E[CLV(T)] = \sum_{t=1}^{\infty} m_{t-1} S(t)$$

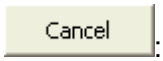
The time slot for which the values of the survival function $S(t)$ are known being limited, we will then talk about long-term value. This explains the term CLTV for customer long-term value.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

ARPA (Average Revenue Per Account): Select the data that correspond to revenues. You must select only one column. If a column header has been selected, check that the "Variable labels" option has been activated.

Time (acquisition/churn): Select in order the two columns: acquisition date and churn date. If a customer has not yet left, the second column of date should be an empty cell. If a column header has been selected, check that the "Variable labels" option has been activated.

Segments: Activate this option if you want to perform the analyses per segment, and then select the data that indicate to which segment each customer belongs. If a column header has been selected, check that the "Variable labels" option has been activated.

Subscription period: Select the subscription period.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections includes a header.

Account names: Activate this option if accounts labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Options tab:

Segments labels: Select the segments labels. This field is only available when the segments option is enabled and at least one of the following options is enabled: discount rate, fixed costs or user defined retention rate.

Discount rate: Activate this option if you want to take into account the discount rate applied to customers. This rate is considered fixed. If the segments option is enabled, select one value per segment and make sure that the lines are in the same order as segments labels.

Fixed costs: Activate this option if you want to deduct certain fixed operating costs from the revenues generated when calculating CLV. Select multiple columns if you have multiple costs to include. If the segments option is enabled, select one value per segment and make sure that the lines are in the same order as segments labels.

Payment:

- **Start of period**: Select this option if payments occur at the beginning of each period.
- **End of period**: Select this option if payments occur at the end of each period.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the revenue variable (ARPA) and a summary of customers actions (number of customers observed, number of customers lost, number of customers for whom we do not have an effective churn date).

CLV: Activate this option to display the average CLV. If the **segments** option is enabled, this value is displayed for each segment.

Customer Lifetime Annalysis: Activate this option to display the customers life cycle analysis tab. If the segment option is enabled, these values are displayed for each segment. You can choose between two options:

- **Summary** : Activate this option if you want to display a resume of your customers life cycle.

- **Full** : Activate this option if you want to display deeper results on your customers life cycle.

Customer long-term value (CLTV) evolution: Activate this option to display, for each period the CLV and the CLTV, which here correspond to the cumulative CLV up to the relevant period.

Segment comparison: If the **segments** option is enabled, activate this option to perform a comparison of the cumulative retention functions of the segments.

Charts tab:

CLV per segment: Activate this option to display the CLV by segment graph as a diagram.

Retention function: Activate this option to display the chart showing the evolution of the cumulative retention function over time.

Probability density: Activate this option to display the chart showing the variation of the probability density as a function of time.

Hazard rate: Activate this option to display the chart showing the hazard rate as a function of time.

Customer long-term value (CLTV) evolution: Activate this option to display the chart showing the CLTV evolution over time.

Segment comparison: If **segments** and **Segment comparison** options are enabled, activate this option to display charts comparing the retention, density and hazard curves for the different segments, if they have been selected.

Note that periods for which censored data have been observed are identified on the chart by a "+".

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics. The number of Observations, the minimum and maximum values, the quartiles, the mean, the standard deviation and the variance (unbiased) are displayed for the revenues variable (ARPA). Also a summary of customers actions (number of customers observed, number of customers lost, number of customers for whom we do not have an effective churn date).

Box plots (or box-and-whisker plots) related to the revenues variable (ARPA) are also displayed (see description).

CLV: The average CLV is displayed. If the segment option is enabled, this value is displayed for each segment.

The **CLV per segment** chart is only available when the segment option is active. This graph displays the CLV for each segment as a bar chart.

Customer Lifetime Annalysis: In this tab are displayed the following results:

- **Summary**:
    - **Period**: Time interval.
    - **Nbr of customers**: Number of customers remaining during the time interval.
    - **Nbr lost**: Number of customers lost during the time interval.
    - **Censored**: Number of customers in the study at the end of the period when it corresponds to the end date of the study.
    - **Effective at risk**: Number of customers that were at risk at the beginning of the interval minus half of the individuals who have been censored during the time interval.
    - **Retention rate**: Proportion of customers remaining during the time interval.
    - **Churn rate**: Proportion of customers lost during the time interval.
- **Full** : Activate this option if you want to display deeper results on your customers life cycle.
- **Cumulative retention function**: Probability that a customer has to continue to be a customer at least until the considered time interval.
- **Cumulative churn function**: Cumulative churn rate until the considered time interval.
- **Probability density**: Estimated density function at the midpoint of the interval.
- **Hazard rate**: Estimated hazard rate function at the midpoint of the interval.

Median retention time: Table displaying the median residual lifetime at the beginning of the experiment, and its standard error. This statistic is one of the key results of the life table analysis as it allows to evaluate the time remaining for half of the customers to "fail".

If the **segments** option is enabled, these values are displayed for each segment.

Segment comparison: You will find in this table a comparison of the cumulative retention functions of the different segments.

This table displays the statistics for three different tests: the Log-rank test, the Wilcoxon test, and the Tarone Ware test. These tests are based on a Chi-square test. The lower the corresponding p-value, the more significant the differences between the segments.

If the p-value obtained by the log-rank test is significant at $\alpha = 5\%$ threshold, then multiple pairwise comparison tests are performed on segments. We perform a Dunn-Sidak test which is a derivative of Bonferroni test and is more efficient in certain situations.

$$\alpha' = 1 - (1 - \alpha)^{1/g}.$$

where g is the number of segments.

- **Segment comparison (charts)**: Depending on the selected options, up to 3 charts with one curve for each group are displayed: cumulative retentio function, probability density

function, hazard rate function.

Example

An example showing how to use the *Customer Long-term Value* is available on the XLSTAT Help Center at:

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

References

Phillip E. Pfeifer, Mark E. Haskins and Robert M. Conroy. Customer Lifetime Value, Customer Profitability, and the Treatment of Acquisition Spending. *Journal of Managerial Issues, Vol. 17, No. 1 (Spring 2005), pp. 11-25*.

Malthouse, Edward C. (2013). Segmentation and Lifetime value Models Using SAS®. *Cary, NC: SAS institute Inc*.

Brookmeyer R. and Crowley J. (1982). A confidence interval for the median survival time. *Biometrics*, **38**, 29-41.

Elandt-Johnson R.C. and Johnson N.L. (1980). Survival Models and Data Analysis. John Wiley & Sons, New York.

Process: moderation and mediation

Use this method in a complementary way to linear regression to deepen the understanding of the phenomenon studied and thus answer the questions "when and how" this effect occurs.

In this section:

Description

Process was popularized by Andrew F. Hayes in 2013. It is a method that is complementary to the PLS-SEM approach and is very popular in the field of marketing as well as in the social and behavioral sciences. Process is decomposed around two main concepts: mediation and moderation.

Mediation analysis is used to test hypotheses clarifying the various intermediate mechanisms by which causal effects arise, while moderation analysis is used to explore questions about the conditions of an effect.

What these two concepts have in common is that they explore the role played by a third variable in the relationship between an explanatory variable X and a response variable Y.

Mediation model

The mediation model assumes that X influences a mediator M which, in turn, influences Y. If X has an effect on Y via M, then the following system summarizes the relationships between X, M and Y : $$ \left\lbrace \begin{aligned} & Y = i_Y + cX +bM +\epsilon_Y \ & M = i_M + aX +\epsilon_M \ \end{aligned} \right. $$ The different parameters of these equations are estimated by the least squares method and make it possible to obtain the indirect effect (which represents the way in which Y is influenced by X through M) and the direct effect (which represents the way in which Y is influenced by X through M) specific to the model: $$ Effect_{Direct}=c $$ $$ Effect_{Indirect}=ab $$ The direct and indirect effects make it possible to conclude on the significance or not of the mediation model.

Moderation model

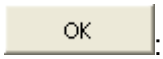
The moderation model assumes that X influences Y more or less strongly depending on a moderator W. If X has an effect on Y modulated by W, then the following equation summarizes the relationships between X, W and Y: $$ Y = i + dX +eW + fInter+\epsilon_Y $$ with $Inter$ representing the interaction variable composed from the variables X and W.

The different parameters of this equation are estimated by the least squares method.

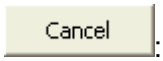
The significance of the moderation is based on the significance or not of the coefficient f associated with the interaction variable.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.


: Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

Y / Dependent variable: Select the response variable you want to model. If a column header has been selected, check that the "Variable labels" option has been activated.

X / Explanatory variable : Select the quantitative explanatory variable in the Excel worksheet. The data selected must be numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

M / Mediator(s) variable(s) : Select the quantitative mediator(s) variable(s) in the Excel worksheet. The data selected must be numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

W / Moderator variable : Select the quantitative moderator variable in the Excel worksheet. The data selected must be numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

V / Moderator variable : Select the quantitative moderator variable in the Excel worksheet. The data selected must be numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

Covariates: Activate this option if you want to add covariates in the model.

Variable labels: Activate this option if the first row of the data selections includes a header.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Model number: Choose the model to use for the calculations (the model is shown on the right side of the dialog box).

Confidence interval (%): Enter the percentage range of the confidence interval to use for calculating the confidence intervals around the parameters. Default value: 95.

Resamplings: Enter the number of samples to generate when boostraping.

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlation matrix: Activate this option to display a view of the correlations between the various variables selected.

Johnson-Neyman chart: Activate this option to display the Johnson-Neyman chart.

Conditional effect chart: Activate this option to display the chart showing the evolution of the conditional effect in the model.

Results

Descriptive statistics: The table of descriptive statistics shows the simple statistics for all the variables selected. The number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the quantitative variables.

Correlation matrix: This table is displayed to give you a view of the correlations between the various variables selected.

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.

- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).

- **R²**: The determination coefficient for the model. This coefficient, which value is between 0 and 1, is defined by:

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ with } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

The R² is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer R² is to 1, the better is the model. The problem with the R² is that it does not take into account the number of variables used to fit the model.

- **Adjusted R²**: The adjusted determination coefficient for the model. The adjusted R² can be negative if the R² is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}$$

The adjusted R² is a correction to the R² which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W - p^*}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

- **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows:

$$MAPE = \frac{100}{W}\sum_{i=1}^{n} w_i \left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

- **DW**: The Durbin-Watson statistic is defined by:

$$DW = \frac{\sum_{i=2}^{n}[(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2}$$

This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- **Cp**: Mallows Cp coefficient is defined by:

$$Cp = \frac{SSE}{\hat{\sigma}} + 2p^* - W$$

where SSE is the sum of the squares of the errors for the model with p explanatory variables and $\hat{\sigma}$ is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to p*, the less the model is biased.

- **AIC**: Akaike's Information Criterion is defined by:

$$AIC = W \ln(\frac{SSE}{W}) + 2p^*$$

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz's Bayesian Criterion is defined by:

$$SBC = W \ln(\frac{SSE}{W}) + \ln(W)p^*$$

This criterion, proposed by Schwarz (1978) is similar to the AIC, and the aim is to minimize it.

- **PC**: Amemiya's Prediction Criterion is defined by:

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}$$

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. The explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable.

The **parameters of the model** table displays the estimate of the parameters, the corresponding standard error, the Student's t, the corresponding probability, as well as the confidence interval.

The **direct effect of X on Y** table displays the estimate of the direct effect of X on Y, the corresponding standard error, the Student's t, the corresponding probability, as well as the confidence interval.

The **conditional direct effect of X on Y** table displays the estimate of the conditional direct effect of X on Y for three values of the moderator (the 16th percentile, the median, and the 84th percentile), the corresponding standard error, the Student's t, the corresponding probability, as well as the confidence interval.

The **indirect effect of X on Y** table displays the estimate of the indirect effect of X on Y, and the corresponding confidence interval and standard deviation obtained by the bootstrap method. If the confidence interval includes 0, then the indirect effect of X on Y in the model is not significant.

The **conditional indirect effect of X on Y** table displays the estimate of the conditional indirect effect of X on Y for three values of the moderator (the 16th percentile, the median, and the 84th percentile), the corresponding standard error, the Student's t, the corresponding probability, as

well as the confidence interval. If the confidence interval includes 0, then for the moderator value in question, the conditional indirect effect of X on Y in the model is not significant.

The **index of moderated mediation** table displays the estimate of the index of moderated mediation, and the corresponding confidence interval and standard deviation obtained by the bootstrap method. If the confidence interval includes 0, the moderated mediation model is not considered significant.

Johnson-Neyman chart: This chart is used to visualize at which value of the moderator the effect becomes significant.

Conditional effect chart: This chart is used to visualize the evolution of the conditional effect in the model for three values of the moderator (the 16th percentile, the median, and the 84th percentile).

Example

A tutorial on how to use the "Process: moderation and mediation" feature is available on the XLSTAT Help Center:

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

References

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.

Conjoint analysis

Design of experiments for conjoint analysis

Use this tool to generate a design for a classical conjoint analysis based on full profiles.

In this section:

Description

The principle of conjoint analysis is to present a set of products (also known as profiles) to the individuals who will rank, rate, or choose some of them.

In an "ideal" analysis, individuals should test all possible products. But in practice it rapidly becomes impossible; the cognitive capacities of a person being limited, and the number of combinations increasing very rapidly with the number of attributes (if one wants to study five attributes with three categories each, it sums up to 243 possible products). We therefore use the methods of experimental design to obtain an acceptable number of profiles to be judged while maintaining good statistical properties.

XLSTAT can generate several unique designs, which is an advantage especially when a large sample of people are interviewed. As the number of different combinations is greater, the analysis designs will be more robust in the analysis of the effects. In addition, including different designs reduces the impact of psychological context and order effects.

XLSTAT-Conjoint includes two different methods of conjoint analysis: the full profile analysis and the choice based conjoint (CBC) analysis.

Full profiles conjoint analysis

The first step in a conjoint analysis requires the selection of a number of factors describing a product. These factors should be qualitative. For example, if one seeks to introduce a new product in a market, we can choose as differentiating factors: the price, the quality, the durability... and for each factor, we must define a number of categories (different prices, different lifetimes...). This first step is crucial and should be carried out in collaboration with experts of the studied market.

Once this first step is done, the goal of a conjoint analysis is to understand the mechanism of choice. Why do people choose one product over another?

To try to answer this question, we will propose a number of products (combining different modalities of the studied factors). We cannot offer all possible products, so we will select products by using design of experiments before presenting them to people who will rate them or rank them.

The full profile method is the oldest method of conjoint analysis; we seek to build an experimental design that includes a limited number of full profiles that each individual interviewed will then rank or rate.

XLSTAT-Conjoint uses fractional factorial designs in order to generate profiles that will then be presented to respondents. When no design is available, XLSTAT-Conjoint uses algorithms to search for D-optimal designs (see description of the module XLSTAT-DOE).

As part of the traditional conjoint analysis, the questionnaires used are based on the rating or ranking of a number of complete profiles.

You have to select the attributes of interest for your product and the categories associated with these attributes. XLSTAT-Conjoint then generates profiles to be ranked / rated by each respondent.

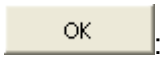
Prohibited combinations

For some experimental designs, categories combinations are not feasible. This may be due to different reasons: equipment, products,... In these cases, it is possible to indicate these prohibited combinations, the generated experimental design will not take them into account.

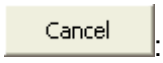
By choosing to add prohibited combinations, the generated plan will necessarily be an optimized plan.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select.

General tab:

Analysis name: Enter the name of the analysis you want to perform.

Factors/Categories table: enter the table containing the names of the factors and their modalities.

Maximum number of profiles: Enter the maximum number of profiles to be presented to the individuals.

Number of responses: Enter the number of expected individuals who will respond to the conjoint analysis.

Number of holdout cases: Enter the number of holdout cases. Holdout cases are cases evaluated by respondents, but which are not subsequently included in the conjoint analysis. These cases are interesting to add to your model because they will allow you to check its validity. They are randomly generated.

- **Randomly mix with other cases**: randomly mixes the holdout cases with the other cases in the experimental design.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Options tab:

Number of designs: Check this option if you want to generate multiple designs.

- Respondents by design: Enter a vector of the size of the number of designs, containing the number of respondents by design.

Prohibited combinations: Check this option if you want to enter prohibited combinations.

Design of experiments:

- D-Optimal design: Select this option to generate a D-Optimal design. This design will correspond exactly to the selected factors (see the chapter on factor effect plans in the help for more details).
- Orthogonal design: Select this option in order to find an orthogonal design close to the parameters entered by the user and present in the XLSTAT database.

Stop conditions: The number of iterations and the convergence criterion to obtain the design can be modified.

Outputs tab:

Optimization summary: Activate this option to display the optimization summary for generating the design.

Burt table: Activate this option to display the Burt table of the experimental design.

Encoded design: Activate this option to display the encoded experimental design in the case of a d-optimal design.

Print individual sheets: Activate this option to print individual sheets, one for each respondent. Each sheet will include all generated profiles. The respondent has to fill the last column of the table with the rates or ranks associated to each generated profile. Two assignment options are available; the fixed option displays the profiles in the same order for all individuals; the random option displays the profiles in random orders (different from one respondent to another).

Include references: Activate this option to include references between the main sheet and the individual sheets. When an individual enters his chosen rating / ranking in the individual sheet, the value is automatically displayed in the main sheet of the analysis.

Prohibited combinations dialog box:

This dialog box allows you to select prohibited combinations. To do this, select the combinations of prohibited modalities in the left-hand part and click on the add button. The combinations will then be displayed in the right- hand part. It is possible to delete one or more selected combinations by clicking on them and then on the remove button. Once you have selected the prohibited combinations, click on the OK button.

Design for conjoint analysis dialog box:

Selection of experimental design: This dialog box lets you select the design of experiments you want to use. A list of fractional factorial designs is presented with their respective distance to the design that was to be generated. If you select a design and you click "Select", then the selected design will appear in your conjoint analysis.

Results

Variable information: This table displays all the information relative to the used factors.

Prohibited combinations information: This table displays all the selected prohibited combinations.

Conjoint analysis design: This table displays the generated profiles. Empty cells associated to each individual respondent are also displayed. If the options "print individual sheets" and "include references" have been activated, then formulas with reference to the individual sheets are included in the empty cells.

Run the analysis: Once you filled in the conjoint design with individuals responses, you can click on the "Run the analysis" button to automatically launch the interface and run a conjoint analysis.

Endoded design ; This table shows the encoded experimental design. This table is only displayed in the case of a d-optimal experimental design.

Burt table: The Burt table is displayed only if the corresponding option is activated in the dialog box. The **3D bar chart** that follows is the graphical visualization of this table.

Optimization details: This table displays the details of the optimization process when a search for a D-optimal design has been selected. This report presents the number of iterations necessary for the realization of the design as well as the D-efficiency or the diagonality. In the case of simple designs, the D-efficiency is displayed. The latter is a relative indicator of efficiency that allows the comparison with other designs of the same size. The goal being to maximize the D-efficiency indicator which lies is within 0 and 1. The diagonality, which is displayed when one wishes to build multiple designs, is an indicator that measures that the confusion between factors and interactions is the minimal one. This closer the diagonality is to 1, the lower the confusion is, which is what we are looking for. The final design selected is bolded in the table.

Individual _Res sheets: When the "Print individual sheets" option is activated, these sheets include the name of the analysis, the individual number and a table associated to the profiles to be rated / ranked. Individual respondents should fill the last column of this table.

Example

An example of full profile based conjoint analysis is available on XLSTAT Help Center:

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

References

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.

Design for choice based conjoint analysis

Use this tool to generate a design of experiments for a Choice-Based Conjoint analysis (CBC).

In this section:

Description

The principle of conjoint analysis is to present a set of products (also known as profiles) to the individuals who will note, class, or choose some of them.

In an "ideal" analysis, individuals should test all possible products. But it is soon impossible; the capacity of each being limited and the number of combinations increases very rapidly with the number of attributes (if one wants to study five attributes with three categories each, that means already 243 possible products). We therefore use the methods of experimental design to obtain a acceptable number of profiles to be judged while maintaining good statistical properties.

XLSTAT can generate several unique designs, which is an advantage especially when a large sample of people are interviewed. As the number of different combinations is greater, the analysis designs will be more robust in the analysis of the effects. In addition, including different designs reduces the impact of psychological context and order effects.

XLSTAT-Conjoint includes two different methods of conjoint analysis: the full profiles analysis and the choice based conjoint (CBC) analysis.

Choice Based Conjoint analysis (CBC)

The principle of choice based conjoint (CBC) analysis is based on choices in a group of profiles. The individual respondent chooses between different products offered instead of rating or ranking products.

The process of CBC is based on comparisons of profiles. These profiles are generated using the same methods as for full profile conjoint analysis. Then, these profiles are put together in many comparison groups (with a fixed size). The individual respondent then chooses the profile that he would select compared to the other profiles included in the comparison.

The statistical process is separated into 2 steps:

- Fractional factorial designs or D-optimal designs are used to generate the profiles.
- Once the profiles have been generated they are allocated in the comparison groups using incomplete block designs.

The first step in a conjoint analysis requires the selection of a number of factors describing a product. These factors should be qualitative. For example, if one seeks to introduce a new product in a market, we can choose as differentiating factors: the price, the quality, the durability... and for each factor, we must define a number of categories (different prices, different lifetimes...). This first step is crucial and should be done together with experts of the studied market.

Once past this first step, the goal of a conjoint analysis is to understand the mechanism for choosing one product over another. Instead of proposing all profiles to the individual respondents and asking to rate or rank them, CBC is based on a choice after a comparison of some of the profiles. Groups of profiles are presented to the individual respondents and they have to indicate which profile they would choose (a no choice option is also available in XLSTAT-Conjoint).

This method combines two designs of experiments, the fractional factorial design to select the profiles to be compared and the incomplete block design to generate the comparisons to be presented. For more details on these methods, please see the screening design chapter of the DOE module help.

XLSTAT-Conjoint enables you to add the no choice option if the individual respondent would not choose any of the proposed profiles.

XLSTAT-Conjoint enables to obtain a global table for CBC analysis but also individual tables for each respondent and each comparison in separated Excel sheets. References are also included so that when a respondent select a profile in an individual sheet, it is directly reported in the main table.

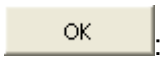
Prohibited combinations

For some experimental designs, categories combinations are not feasible. This may be due to different reasons: equipment, products,... In these cases, it is possible to indicate these prohibited combinations, the generated experimental design will not take them into account.

By choosing to add prohibited combinations, the generated plan will necessarily be an optimized plan.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select.

General tab:

Analysis name: Enter the name of the analysis you want to perform.

Factors/Categories table: enter the table containing the names of the factors and their modalities.

Maximum number of profiles: Enter the maximum number of profiles to be presented to the individuals.

Number of responses: Enter the number of expected individuals who will respond to the conjoint analysis.

Maximum number of comparisons: Enter the maximum number of comparison to be presented to the individual respondents. This number has to be greater than the number of profiles.

Number of profiles per comparison: Enter the number of profiles per comparison.

Number of holdout cases: Enter the number of holdout cases. Holdout cases are cases evaluated by respondents, but which are not subsequently included in the conjoint analysis. These cases are interesting to add to your model because they will allow you to check its validity. They are randomly generated.

- **Randomly mix with other cases**: randomly mixes the holdout cases with the other cases in the experimental design.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Options tab:

Number of designs: Check this option if you want to generate multiple designs.

- Respondents by design: Enter a vector of the size of the number of designs, containing the number of respondents by design.
- Comparisons by design: Enter a vector of the size of the number of designs, containing the number of combinations by design.

Prohibited combinations: Check this option if you want to enter prohibited combinations.

Design of experiments:

- D-Optimal design: Select this option to generate a D-Optimal design. This design will correspond exactly to the selected factors (see the chapter on factor effect plans in the help for more details).
- Orthogonal design: Select this option in order to find an orthogonal design close to the parameters entered by the user and present in the XLSTAT database.

Stop conditions: The number of iterations and the convergence criterion to obtain the design can be modified.

Outputs tab:

Optimization summary: Activate this option to display the optimization summary for generating the design.

Burt table: Activate this option to display the Burt table of the experimental design.

Encoded design: Activate this option to display the encoded experimental design in the case of a d-optimal design.

Print individual sheets: Activate this option to print individual sheets,one for each respondent. Each sheet will include a table for each comparison. The respondent has to enter the code associated to the profile he would choose in the box at the bottom of each table. Two assignment options are available; the fixed option displays the comparisons in the same order for all individuals; the random option displays the comparisons in random orders (different from one respondent to another).

Include references: Activate this option to include references between the main sheet and the individual sheets. When an individual enters his chosen code in the individual sheet, the result is automatically displayed in the main sheet of the analysis.

Include the no choice option: Activate this option to include a no choice option for each comparison in the individual sheets.

Prohibited combinations dialog box:

This dialog box allows you to select prohibited combinations. To do this, select the combinations of prohibited modalities in the left-hand part and click on the add button. The combinations will then be displayed in the right- hand part. It is possible to delete one or more selected combinations by clicking on them and then on the remove button. Once you have selected the prohibited combinations, click on the OK button.

Design for conjoint analysis dialog box:

Selection of experimental design: This dialog box lets you select the design of experiment you want to use. Thus, a list of fractional factorial designs is presented with their respective distance to the design that was to be generated. If you select a design and you click Select, then the selected design will appear in your conjoint analysis.

Results

Variable information: This table displays all the information relative to the used factors.

Prohibited combinations information: This table displays all the selected prohibited combinations.

Profiles: This table displays the generated profiles using the design of experiments tool.

Conjoint analysis design: This table displays the comparisons presented to the respondent. Each row is associated to a comparison of profiles. The numbers in the rows are associated to the profiles numbers in the profiles tables. Empty cells associated to each individual respondent are also displayed. Respondent have to enter the code associated to the choice made (1 to number of profiles per comparisons; or 0 if the no choice option is selected).

Run the analysis: Once you filled in the conjoint design with individuals responses, you can click on the "Run the analysis" button to automatically launch the interface and run a conjoint analysis based on choice.

Endoded design ; This table shows the encoded experimental design. This table is only displayed in the case of a d-optimal experimental design.

Burt table: The Burt table is displayed only if the corresponding option is activated in the dialog box. The **3D bar chart** that follows is the graphical visualization of this table.

Optimization details: This table displays the details of the optimization process when a search for a D-optimal design has been selected. This report presents the number of iterations necessary for the realization of the design as well as the D-efficiency or the diagonality. In the case of simple designs, the D-efficiency is displayed. The latter is a relative indicator of efficiency that allows the comparison with other designs of the same size. The goal being to maximize the D-efficiency indicator which lies is within 0 and 1. The diagonality, which is displayed when one wishes to build multiple designs, is an indicator that measures that the confusion between factors and interactions is the minimal one. This closer the diagonality is to 1, the lower the confusion is, which is what we are looking for. The final design selected is bolded in the table.

Individual _Res sheets: When the "Print individual sheets" option is activated, these sheets include the name of the analysis, the individual number and tables associated to the comparisons with the profiles to be compared. Individual respondents should enter the code associated to their choice in the bottom right of each table.

Example

An example of choice based conjoint (CBC) analysis is available on XLSTAT Help Center:

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

References

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.

Conjoint analysis

Use this tool to run a Full Profile Conjoint analysis. This tool is included in the XLSTAT-Conjoint module; it must be applied on design of experiments for conjoint analysis generated with XLSTAT-Conjoint.

In this section:

Description

Conjoint analysis is a comprehensive method for the analysis of new products in a competitive environment.

This tool allows you to carry out the step of analyzing the results obtained after the collection of responses from a sample of people. It is the fourth step of the analysis, once the attributes have been defined, the design has been generated and the individual responses have been collected.

Full profile conjoint analysis is based on ratings or rankings of profiles representing products with different characteristics. These products have been generated using a design of experiments and can be real or virtual.

The analysis is done using two statistical methods:

- Analysis of variance based on ordinary least squares (OLS).

- Monotone analysis of variance (Kruskal, 1964) that uses monotonic transformations of the responses to better adjust the analysis of variance (MONANOVA).

Both approaches are described in detail in the chapters "Analysis of variance" and "Monotone regression (MONANOVA)" of the help of XLSTAT.

Conjoint analysis therefore provides for each individual what is called partial utilities associated with each category of the variables. These utilities provide a rough idea of the impact of each modality on the process of choosing a product.

In addition to utilities, conjoint analysis provides an importance associated with each variable. It shows how each variable in the selection process associated with each individual is important.

The full profile conjoint analysis details the results for each individual separately, which preserves the heterogeneity of the results.

XLSTAT-Conjoint also proposes to make classifications on the individuals. Using the utilities, XLSTAT-Conjoint will obtain classes of individuals that can be analyzed and be useful for further research. Classification methods used in XLSTAT-Conjoint are the agglomerative hierarchical classification (see the chapter on this subject in the help of XLSTAT) and the k-means method (see the chapter on this subject in the help of XLSTAT).

Type of data

XLSTAT-Conjoint offers two types of input data for the conjoint analysis: rankings and ratings. The type of data must be indicated because the treatment used is slightly different.

Indeed, with rankings, the best profile will have the lowest value, whereas with a rating, it will have the highest value.

If the ranking option is selected, XLSTAT-Conjoint transforms the answers in order to reverse this arrangement and so that utilities can be interpreted easily.

Interactions

By interaction is meant an artificial factor (not measured) which reflects the interaction between at least two measured factors. For example, if we carry out treatment on a plant, and tests are carried out under two different light intensities, we will be able to include in the model an interaction factor treatment*light which will be used to identify a possible interaction between the two factors. If there is an interaction between the two factors, we will observe a significantly larger effect on the plants when the light is strong and the treatment is of type 2 while the effect is average for weak light, treatment 2 and strong light, treatment 1 combinations.

To make a parallel with linear regression, the interactions are equivalent to the products between the continuous explanatory values although here obtaining interactions requires nothing more than simple multiplication between two variables. However, the notation used to represent the interaction between factor A and factor B is A*B.

The interactions to be used in the model can be easily defined in XLSTAT.

Constraints

During the calculations, each factor is broken down into a sub-matrix containing as many columns as there are categories in the factor. Typically, this is a full disjunctive table. Nevertheless, the breakdown poses a problem: if there are g categories, the rank of this sub-matrix is not g but g-1. This leads to the requirement to delete one of the columns of the sub-matrix and possibly to transform the other columns. Several strategies are available depending on the interpretation we want to make afterwards:

1) **a1=0**: the parameter for the first category is null. This choice allows us force the effect of the first category as a standard. In this case, the constant of the model is equal to the mean of the

dependent variable for group 1.

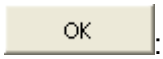
2) **an=0**: the parameter for the last category is null. This choice allows us force the effect of the last category as a standard. In this case, the constant of the model is equal to the mean of the dependent variable for group g.

3) **Sum (ai) = 0**: the sum of the parameters is null. This choice forces the constant of the model to be equal to the mean of the dependent variable when the design is balanced.

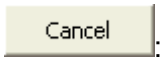
Note: even if the choice of constraint influences the values of the parameters, it has no effect on the predicted values and on the different fitting statistics.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Automatically load data  : In order to run a conjoint analysis you need to load two data tables: a table with the individual responses and a table containing the different profiles (products rated by respondents). If the design experiment has been generated with XLSTAT , then you can load the two tables automatically. To do that, click on the "magic stick" button and select any cell in the sheet containing the design generated by XLSTAT. In order to correctly load your data it is very important that you did not manually modified the sheet containing the design generated by XLSTAT (no add of rows or columns,…). You can also load your data manually.

Responses: Select the responses that have been given by the respondents. If headers have been selected, please check the option "Variable labels" is enabled. This selection corresponds to the right part of the conjoint analysis design table generated with the "design of conjoint analysis" tool of XLSTAT- Conjoint.

Response type: select the type of response given by the respondents (ratings or rankings).

Profiles: Select the profiles that have been generated. If headers have been selected, please check the option "Variable labels" is enabled. This selection corresponds to the right part of the conjoint analysis design table generated with the "design of conjoint analysis" tool of XLSTAT-Conjoint. Do not select the first column of the table.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selections (data, other group) contains a label.

Profiles weights: Activate this option if profiles weights are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection.

Holdout cases: Enter the holdout cases column. Holdout cases are cases evaluated by respondents, but which are not subsequently included in the conjoint analysis. These cases are interesting to add to your model because they will allow you to check its validity. In the case of the conjoint analysis, a measure of Kendall's Tau per respondent will be calculated for these cases specifically.

- **Randomly mix with other cases**: randomly mixes the holdout cases with the other cases in the experimental design.

Options tab:

Method: Select the method to be used for estimation.

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Tolerance: Activate this option to prevent the OLS regression calculation algorithm taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions.

Default value: 95.

Constraints: Details on the various options are available in the description section.

- **a1 = 0**: Choose this option so that the parameter of the first category of each factor is set to 0.
- **an = 0**: Choose this option so that the parameter of the last category of each factor is set to 0.
- **Sum (ai) = 0**: for each factor, the sum of the parameters associated with the various categories is set to 0.

Segmentation: Activate this option if you want XLSTAT-Conjoint to apply an individuals based clustering method on the on the partial utilities. Two methods are available: agglomerative hierarchical classification and k-means classification.

- **Number of classes**: Enter the number of classes to be created by the algorithm for the k-means.
- **Truncation**: Activate this option if you want XLSTAT to **automatically** define the truncation level, and therefore the number of classes to retain, or if you want to define the **number of classes** to create, or the **level** at which the dendrogram is to be truncated.

Stop conditions: the number of iterations and the convergence criterion for the MONANOVA algorithm can be modified.

Missing data tab:

Remove responses of respondents: Activate this option to remove all the responses of respondents having one or more missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean of ratings of respondents for the corresponding profile.
- **Nearest neighbor**: Activate this option to estimate the missing data of a respondent by using the ratings of the nearest respondent for the corresponding profile.

Outputs tab:

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Type III analysis: Activate this option to display the type III analysis of variance table.

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

Charts tab:

Utilities chart: Activate this option to display chart of average utilities.

Importances chart: Activate this option to display chart of average importances.

Regression charts: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.

Transformation plot: Activate this option to display the monotone transformation of the responses plot.

Dendrogram: Activate this option to display the dendrogram.

- **Horizontal**: Choose this option to display a horizontal dendrogram.
- **Vertical**: Choose this option to display a vertical dendrogram.
- **Full**: Activate this option to display the full dendrogram (all objects are represented).
- **Truncated**: Activate this option to display the truncated dendrogram (the dendrogram starts at the level of the truncation).
- **Labels**: Activate this option to display object labels (full dendrogram) or classes (truncated dendrogram) on the dendrogram.
- **Colors**: Activate this option to use colors to represent the different groups on the full dendrogram.

Results

Variable information: This table displays all the information relative to the used factors.

Utilities (individual data): This table displays utilities associated to each category of the factors for each respondent.

Standard deviations table: This table displays the standard deviation for each utility and each respondent together with the model error. It is useful to apply the RFC-Bolse Approach for market simulation (see the conjoint analysis simulation chapter).

Utilities (descriptive statistics): This table displays minimum, maximum, mean and standard error of the partial utilities associated to each category of the factors.

Importance (individual data): This table displays importance for each factor of the analysis for each respondent.

Importance (descriptive statistics): This table displays minimum, maximum, mean and standard error of the importance for each factor of the analysis.

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.
- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).
- R^2: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2}, \text{ where } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} w_i y_i$$

The R^2 is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer R^2 is to 1, the better is the model. The problem with the R^2 is that it does not take into account the number of variables used to fit the model.

- **Adjusted** R^2: The adjusted determination coefficient for the model. The adjusted R^2 can be negative if the R^2 is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}$$

The adjusted R^2 is a correction to the R^2 which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W - p^*}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.
- **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows::

$$MAPE = \frac{100}{W}\sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- **DW**: The Durbin-Watson statistic is defined by:

$$DW = \frac{\sum_{i=2}^{n}[(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}$$

This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- Cp: Mallows Cp coefficient is defined by:

$$Cp = \frac{SCE}{\hat{\sigma}} + 2p^* - W$$

where SSE is the sum of the squares of the errors for the model with p explanatory variables and $\hat{\sigma}$ is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to $p*$, the less the model is biased.

- **AIC**: Akaike's Information Criterion is defined by:

$$AIC = W \ln\left(\frac{SCE}{W}\right) + 2p^*$$

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz's Bayesian Criterion is defined by:

$$SBC = W \ln\left(\frac{SCE}{W}\right) + \ln(W)p^*$$

This criterion, proposed by Schwarz (1978) is similar to the AIC and, like this, the aim is to minimize it.

- **PC**: Amemiya's Prediction Criterion is defined by:

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}$$

This criterion, proposed by Amemiya (1980) is used, like the adjusted R^2 to take account of the parsimony of the model.

- **Press RMSE**: Press' statistic is only displayed if the corresponding option has been activated in the dialog box. It is defined by:

$$Press = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i(-i)})^2$$

where $\hat{y}_{i(-i)}$is the prediction for observation i when the latter is not used for estimating parameters. We then get:

$$\text{Press RMCE } = \sqrt{\frac{\text{Press}}{W - p^*}}$$

Press's RMSE can then be compared to the RMSE. A large difference between the two shows that the model is sensitive to the presence or absence of certain observations in the model.

- **Iteration**: Number of iteration until convergence of the ALS algorithm.

Goodness of fit coefficients (MONANOVA): In this table are shown the statistics for the fit of the regression model specific to the case of MONANOVA. These statistics are the Wilks' lambda, the Pillai's trace, the trace of Hotelling-Lawlet and the largest root of Roy. For more details on these statistics, see the help on the conditional logit model.

If the Type I/II/III SS (SS: Sum of Squares) option is activated, the corresponding tables are displayed.

The table of **Type I SS** values is used to visualize the influence that progressively adding explanatory variables has on the fitting of the model, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. The sums of squares in the Type I table always add up to the model SS. Note: the order in which the variables are selected in the model influences the values obtained.

The table of **Type II SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. Type II SS are not recommended in unbalanced designs but we display them as some users might need them. It is identical to Type III for balanced designs.

The table of **Type III SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. While Type II SS depends on the number of observations per cell (cell means combination of categories of the factors), Type III does not and is therefore preferred.

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

The **parameters of the model** table displays the estimate of the parameters, the corresponding standard error, the Student's t, the corresponding probability, as well as the confidence interval

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the observed value of the dependent variable, the transformed value of the dependant variable, the model's prediction, the residuals, and the confidence intervals. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the random values being larger.

The **chart** which follows shows the transformation of the dependant variable.

If a segmentation method has been applied, the following results are displayed:

Class centroids: This table shows the class centroids for the various descriptors.

Proximity matrix: This table displays the proximities between the respondents calculated on their associated utilities.

Dendrograms: The full dendrogram displays the progressive clustering of objects. If truncation has been requested, a broken line marks the level the truncation has been carried out. The truncated dendrogram shows the classes after truncation.

Results by class: The descriptive statistics for the classes (number of objects, sum of weights, within-class variance, minimum distance to the centroid, maximum distance to the centroid, mean distance to the centroid) are displayed in the first part of the table. The second part shows the objects.

Results by object: This table shows the assignment class for each object in the initial object order.

Example

An example conjoint analysis is available on XLSTAT Help Center:

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

References

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.

Choice based conjoint analysis

Use this tool to run a Choice-Based Conjoint analysis (CBC). This tool is included in the XLSTAT-Conjoint module; it must be applied on design of experiments for choice based conjoint analysis generated with XLSTAT-Conjoint.

In this section:

Description

Conjoint analysis is a comprehensive method for the analysis of new products in a competitive environment.

This tool allows you to carry out the step of analyzing the results obtained after the collection of responses from a sample of people. It is the fourth step of the analysis, once the attributes have been defined, the design has been generated and the individual responses have been collected.

In the case of CBC models, individuals have to choose between selections of profiles. Thus, a number of choices are given to all individuals (we will select a product from a number of products generated).

Analysis of these choices is made using:

- A multinomial logit model based on a specific conditional logit model. For more details see the help on the conditional logit model. In this case, we obtain aggregate utilities, that is to say, one utility for each category of each variable associated with all the individuals. It is impossible to make classifications based on the individuals.
- A hierarchical Bayes algorithm which gives individual results. Parameters are estimated at the individual level using an iterative method (Gibbs sampling) taking into account each individual's choice but also the global distribution of the choices. The obtained individual utilities will give better market simulation as the classical CBC algorithm.

XLSTAT-Conjoint proposes to include a segmentation variable when using the classical CBC algorithm that will build separate models for each group defined by the variable.. When CBC/HB is used, since individual utilities are obtained, you can apply a clustering method on the individuals.

In addition to utilities, conjoint analysis provides the importance associated with each variable.

Interactions

By interaction is meant an artificial factor (not measured) which reflects the interaction between at least two measured factors. For example, if we carry out treatment on a plant, and tests are carried out under two different light intensities, we will be able to include in the model an interaction factor treatment*light which will be used to identify a possible interaction between the two factors. If there is an interaction between the two factors, we will observe a significantly larger effect on the plants when the light is strong and the treatment is of type 2 while the effect is average for weak light, treatment 2 and strong light, treatment 1 combinations.

To make a parallel with linear regression, the interactions are equivalent to the products between the continuous explanatory values although here obtaining interactions requires nothing more than simple multiplication between two variables. However, the notation used to represent the interaction between factor A and factor B is A*B.

The interactions to be used in the model can be easily defined in XLSTAT.

Constraints

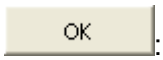
During the calculations, each factor is broken down into a sub-matrix containing as many columns as there are categories in the factor. Typically, this is a full disjunctive table. Nevertheless, the breakdown poses a problem: if there are g categories, the rank of this sub-matrix is not g but g-1. This leads to the requirement to delete one of the columns of the sub-matrix and possibly to transform the other columns. Several strategies are available depending on the interpretation we want to make afterwards:

- **a1=0**: the parameter for the first category is null. This choice allows us force the effect of the first category as a standard.
- **an=0**: the parameter for the last category is null. This choice allows us force the effect of the last category as a standard.
- **Sum (ai) = 0**: the sum of the parameters is null.

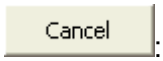
Note: even if the choice of constraint influences the values of the parameters, it has no effect on the predicted values and on the different fitting statistics.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Automatically load data  : In order to run a choice based conjoint analysis you need to load three data tables: a table with individual responses, a table containing the different choices (comparisons) and a table containing the different profiles (products respondents have to choose). If the design experiment has been generated with XLSTAT , then you can load the three tables automatically. To do that, click on the "magic stick" button and select any cell in the sheet containing the design generated by XLSTAT. In order to correctly load your data it is very important that you did not manually modified the sheet containing the design generated by XLSTAT (no add of rows or columns,…). You can also load your data manually.

Responses: Select the responses that have been given by respondents. If headers have been selected, please check the option "Variable labels" is enabled. This selection corresponds to the right part of the conjoint analysis design table generated with the "design of choice based conjoint analysis" tool of XLSTAT-Conjoint.

Choice table: Select the choices that have been presented to the respondents. If headers have been selected, please check the option "Variable labels" is enabled. This selection corresponds to the left part of the conjoint analysis design table generated with the "design of choice based conjoint analysis" tool of XLSTAT-Conjoint. Do not select the first column of the table.

Profiles: Select the profiles that have been generated. If headers have been selected, please check the option "Variable labels" is enabled. This selection corresponds to profiles table generated with the "design of choice based conjoint analysis" tool of XLSTAT-Conjoint. Do not select the first column of the table.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selections (data, other group) contains a label.

Group variable: Activate this option then select a column containing the group identifiers. If a header has been selected, check that the "Variable labels" option has been activated.

Response weights: Activate this option if response weights are available. Then select the corresponding data. If the"Variable labels" option is activated you need to include a header in the selection.

Holdout cases: enter the holdout cases column. Holdout cases are cases evaluated by respondents, but which are not subsequently included in the conjoint analysis. These cases are interesting to add to your model because they will allow you to check its validity. In the case of the multinomial logit model, a measure of rlh will be calculated for these cases specifically. In the case of the hierarchical Bayes model, a measure of rlh per respondent will be calculated for these cases.

Options tab:

Method: Select the method to be used for estimation.

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Tolerance: Activate this option to prevent the OLS regression calculation algorithm taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95%.

Constraints: Details on the various options are available in the description section.

a1 = 0: Choose this option so that the parameter of the first category of each factor is set to 0.

an = 0: Choose this option so that the parameter of the last category of each factor is set to 0.

Sum (ai) = 0: for each factor, the sum of the parameters associated with the various categories is set to 0.

Bayesian options (only using CBC/HB algorithm): the number of iterations for the burn-in period and the maximal time for the hierarchical Bayes algorithm can be modified.

Segmentation (only using CBC/HB algorithm): Activate this option if you want to apply an individual based clustering method on the partial utilities. Two methods are available: agglomerative hierarchical classification and k-means classification.

- **Number of classes**: Enter the number of classes to be created by the algorithm for the k-means.
- **Truncation**: Activate this option if you want XLSTAT to automatically define the truncation level, and therefore the number of classes to retain, or if you want to define the number of classes to create, or the level at which the dendrogram is to be truncated.

Stop conditions: the number of iterations and the convergence criterion until convergence of the Newton-Raphson algorithm can be modified.

Missing data tab:

Remove responses of respondent: Activate this option to remove all the responses of the respondent with missing data.

Outputs tab:

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Type III analysis: Activate this option to display the type III analysis of variance table.

Model coefficients: Activate this option to display the model's coefficients also called aggregated utilities.

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Predictions and residuals: Activate this option to display the predictions and residuals obtained with the aggregated utilities.

Observation details (only using CBC/HB algorithm): activate this option to display the characteristics of the posterior distribution for each individual when using CBC/HB algorithm.

Charts tab:

Utilities chart: Activate this option to display chart of average utilities.

Importances chart: Activate this option to display chart of average importances.

Dendrogram (only using CBC/HB algorithm): Activate this option to display the dendrogram.

- **Horizontal**: Choose this option to display a horizontal dendrogram.
- **Vertical**: Choose this option to display a vertical dendrogram.
- **Full**: Activate this option to display the full dendrogram (all objects are represented).
- **Truncated**: Activate this option to display the truncated dendrogram (the dendrogram starts at the level of the truncation).
- **Labels**: Activate this option to display object labels (full dendrogram) or classes (truncated dendrogram) on the dendrogram.
- **Colors**: Activate this option to use colors to represent the different groups on the full dendrogram.

Results

Variable information: This table displays all the information relative to the used factors.

XLSTAT displays a large number tables and charts to help in analyzing and interpreting the results.

Utilities: This table displays utilities associated to each category of the factors with their respective standard error.

Importance: This table displays importance for each factor of the analysis.

Goodness of fit coefficients: This table displays a series of statistics for the independent model (corresponding to the case where the linear combination of explanatory variables reduces to a constant) and for the adjusted model.

- **Observations**: The total number of observations taken into account (sum of the weights of the observations);
- **Sum of weights**: The total number of observations taken into account (sum of the weights of the observations multiplied by the weights in the regression);
- **DF**: Degrees of freedom;
- **-2 Log(Like.)** : The logarithm of the likelihood function associated with the model;
- R^2**(McFadden)**: Coefficient, like the R^2, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model;
- R^2**(Cox and Snell)**: Coefficient, like the R^2, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model raised to the power $\frac{2}{S_w}$, where S_w is the sum of weights.

- R^2**(Nagelkerke)**: Coefficient, like the R^2, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to ratio of the R^2 of Cox and Snell, divided by 1 minus the likelihood of the independent model raised to the power $\frac{2}{S_w}$;
- **AIC**: Akaike's Information Criterion;
- **SBC**: Schwarz's Bayesian Criterion.
- **Iteration**: Number of iteration to reach convergence.
- **rlh**: root likelihood. This value varies between 0 and 1, the value of 1 being a perfect fit.
- **rlh by individual**: The RLH (Root LikeliHood) value is an index from 0 to 1. The higher the RLH value of a respondent, the more consistently the respondent answered the choice questions."

Goodness of fit indexes (conditional logit): In this table are shown the goodness of fit statistics specific to the case of the conditional logit model. For more details on these statistics, see the description part of this help.

Test of the null hypothesis H0: Y=p0: The H_0 hypothesis corresponds to the independent model which gives probability p_0 whatever the values of the explanatory variables. We seek to check if the adjusted model is significantly more powerful than this model. Three tests are available: the likelihood ratio test (-2 Log(Like.)), the Score test and the Wald test. The three statistics follow a χ^2 distribution whose degrees of freedom are shown.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can easily be seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

Example

An example of choice based conjoint (CBC) analysis is available on XLSTAT Help Center:

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

References

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.

Market generator

Use this tool to generate a market with different products which will be used to simulate market share of the different products with the simulation tool.

In this section:

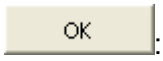
Description

Results of a conjoint analysis or a conjoint analysis based on choice can be used to simulate the market shares of different products with the simulation tool. To do that, it is necessary to generate a table describing the different products of the market to simulate. The XLSTAT "Market generator" tool allows to create this table from the table named "variable information" generated with a conjoint analysis.

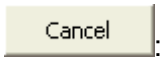
Once this information is entered into the dialog box, just click OK, and for each attribute of each product, you will be asked to choose the category to add. When an entire product has been defined, you can either continue with the next product or stop building the table and obtain a partial market table.

Dialog box

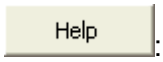
The dialog box contains one tab that correspond to the options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

Variable information: Select the table named "variable information" (with variable labels) generated with results of a conjoint analysis or a conjoint analysis based on choice.

Number of products: Enter the maximum number of products you want to generate.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Conjoint analysis simulation tool

Use this tool to run market simulations based on the results of a conjoint analysis (full profile or choice-based) obtained with XLSTAT-Conjoint.

In this section:

Description

Conjoint analysis is a comprehensive method for the analysis of new products in a competitive environment.

Once the analysis has been performed, the major advantage of conjoint analysis is its ability to perform market simulations using the obtained utilities. The products included in the market do not have to be part of the tested products.

Outputs from conjoint analysis include utilities which can be partial (associated to each individual in full profile conjoint analysis) or aggregate (associated to all the individuals in CBC). These utilities allow computing a global utility associated to any product that you want to include in your simulated market.

Four estimation methods are proposed in XLSTAT-Conjoint: first choice, logit, Bradley-Terry-Luce and randomized first choice. These methods are described bellow.

The obtained market shares can then be analyzed to assess the possible introduction of a new product on the market. The results of these simulations are nevertheless dependent on the knowledge of the real market and the fact that all important factors associated with each product in the conjoint analysis have been taken into account.

XLSTAT-Conjoint can also add weights to the categories of the factors or to the individuals.

XLSTAT-Conjoint can also take into account groups of individuals when a group variable (segmentation) is available. It can be obtained, for example, with the segmentation tool associated with the conjoint analysis.

Data type

XLSTAT-Conjoint proposes two models for conjoint analysis. In a full profile analysis, a constant is associated to the utilities and there are as many utilities as individuals in the study. You have to select all the utilities and their constant (without the column with the names of the categories). In the case of CBC, there is no constant and you have to select one column of utilities without the labels associated to the name of the categories.

In XLSTAT-Conjoint, you have to entirely select the variable information table provided by the conjoint analysis tool. On the other hand, the market to be simulated can be generated with the "Market generator" XLSTAT function by using the table containing variables information generated in a conjoint analysis.

Simulation methods

XLSTAT-Conjoint offers four methods for simulation of market share.

The first step consists of calculating the global utility associated with each new product. Thus, for a CBC analysis for analyzing men's shoes with three factors: the price (50 dollars, 100 dollars, 150 dollars), their finishing (canvas, leather, suede) and the color (brown, black). We have a table with 8 partial utilities rows and one column.

We want to simulate a market with a black leather shoe with price equal USD 100. The utility of this product is: $U_{P1} = U_{Prix-100} + U_{F-Cuir} + U_{C-Noir}$

We calculate the utility for each product in the market and we seek the probability of choosing this product using different estimation methods:

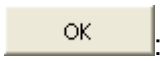
- First choice: it is the most basic; you select the product with maximum utility with a probability of 1.
- Logit: this method is based on the exponential function to find the probability, it is more accurate than the method first choice and it is generally preferred. It has the disadvantage of the IIA assumption (assumption of independence of irrelevant alternatives). It is calculated for the product $P1$: $P_{P1} = \frac{\exp(U_{P1}\beta)}{sum_i \exp(U_{Pi}\beta)}$ with $\beta = 1$ or 2.
- Bradley-Terry-Luce is a method close to the logit method without using the exponential function. It always involves the assumption of IIA and demands positive utilities (if $\beta = 1$). It is calculated for the product $P1$: $P_{P1} = \frac{U_{P1}^{\beta}}{\sum_i U_{Pi}^{\beta}}$ with $\beta = 1$ or 2.
- Randomized first choice: it is a method midway between logit and First Choice. It has the advantage of not assuming the IIA assumption and is based on a simple principle: it generates a large number of numbers from a Gumbel distribution and creates a new set of utilities using the initial utilities adding the numbers generated. For each set of utilities created, the first choice method is used to select one of the products. So we will accept slight variations around the calculated values of the utilities. This method is the most advanced but also more suited to the case of conjoint analysis.
- RFC-Bolse: In the case of profile-based conjoint analysis, the Randomized First Choice BOLSE (RFC-BOLSE) was introduced to overcome the problems of the RFC method. Indeed, RFC is based on a Gumbel law that do not fit the full profile method. This

approach is based on the same principle as Randomized First Choice but it uses a different distribution function to generate the simulated numbers. The RFC model adds unique random error (variation) to the part-worths and computes market shares using the First Choice rule. The centered normal distribution is used with standard error equal to the standard error of the parameters of the regression model and a global error term associated to the entire model. For each set of utilities created, the first choice method is used to select one of the products. So we will accept slight variations around the calculated values of the utilities. This method is the most advanced but also more suited to the case of profile based conjoint analysis.

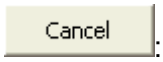
When more than one column of utilities (with a conjoint analysis with full profiles) are selected, XLSTAT-Conjoint uses the mean of the probabilities.

Dialog box

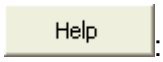
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Automatically load data  : In order to run a conjoint analysis simulation, you need to load three data tables: the table with utilities, the table with information on variables generated with a conjoint analysis made with XLSTAT and the table containing the market to simulate. If the conjoint analysis has been generated with XLSTAT and if you generated the market to simulate with the "Market generator" XLSTAT function, then you can load the three tables automatically. To do that, click on the "magic stick" button and select any cell of the sheet containing conjoint analysis results, then select any cell of the sheet containing the generated market by XLSTAT. In order to correctly load your data it is very important that you did not manually modified the sheets containing the results and the market generated by XLSTAT (no add of rows or columns, …). You can also load your data manually.

Utilities table: Select the utilities obtained with XLSTAT-Conjoint. If headers have been selected, please check the option "Variable labels" is enabled. Do not select the name of the categories.

Variables information: Select the variable information table generated with XLSTAT-Conjoint. If headers have been selected, please check the option "Variable labels" is enabled.

Model: Choose the type of conjoint analysis that you used (full profile or CBC).

Simulated market: Select the market to be simulated. The products will be distributed in a table with a product per line and a variable per column. If headers have been selected, please check the option "Variable labels" is enabled.

Method: Choose the method to use to compute market shares.

Product ID: Activate this option if products ID are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selections (data, other group) contains a label.

Categories weights: Activate this option if categories weights are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection.

Group variable: Activate this option then select a column containing the group identifiers. If a header has been selected, check that the "Variable labels" option has been activated.

Response weights: Activate this option if response weights are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection.

Options tab:

Interactions / Level: Activate this option if interactions were selected in the conjoint analysis. Then, enter the maximum level of interaction (value between 1 and 3).

Number of simulations: Enter the number of simulations to be generated with the "randomized first choice" option.

Standard deviations table: Select the table of standard deviations obtained from the full profile conjoint analysis. If headers have been selected, please check the option "Variable labels" is enabled.

Charts tab:

Market share plot: Activate this option to display market share plots:

- **Pie charts**: Activate this option to display market share pie charts.
- **Compare to the total sample**: If groups have been selected, activate this option to compare the market shares of sub-samples with those of the complete sample.

Results

Variable information: This table displays the summary of the information on the selected factors.

Simulated market: This table displays the products used to perform the simulation. Usually, first products correspond to product currently in the market place while the last product correspond to a new product you want to test on the existing market.

Run again the analysis: In order to better understand influences of the different characteristics of products on market shares, you can modify categories of last product. Once you modified the categories, by clicking on the "Run again the analysis" button, market shares and associated charts will be automatically updated according the new categories of the last product.

Market shares: This table displays the obtained market shares. If groups have been selected, the first column is associated with the global market and the following columns are associated with each group

Market share plots: The first pie chart is associated to the global market. If groups have been selected, the following diagrams are associated with the different groups. If the option "compare to the total sample" is selected, the plots are superimposed; in the background the global market shares are displayed and in front, market shares associated to the group of individuals studied are shown.

Utilities / Market shares: This table, which appears only if no groups are selected, displays products utilities, market shares as well as standard deviations (when possible) associated with each product from the simulated market.

Market shares (individual): This table, which appears only if no groups are selected and when full profile conjoint analysis is selected, displays market shares obtained for each individual.

Example

An example of conjoint analysis is available on XLSTAT Help Center:

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

An example of choice based conjoint (CBC) analysis is available on XLSTAT Help Center:

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

References

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.

Design for MaxDiff

Use this tool to generate a design of experiments for MaxDiff analysis (best- worst model).

In this section:

Description

MaxDiff or Maximum Difference Scaling is a method introduced by Jordan Louvière (1991) that allows obtaining importance of attributes. Attributes are presented to a respondent who must choose the best and worst attributes (most important / least important).

Two steps are needed to apply that method:

- First, a design must be generated so that each attribute is presented with other attributes an equal number of times.
- Then, once the respondent has selected for each choice the best and worst attribute, a model is applied in order to obtain the importance of each attribute. A Hierarchical Bayes model is applied to obtain individual values of the importance.

To obtain the MaxDiff design, design of experiments is used. An incomplete block design is used to generate the choices to be presented. For more details on these methods, please see the DOE for sensory analysis chapter.

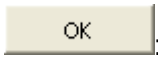
The number of comparisons and the number of attributes per comparison should be chosen depending on the number of attributes. Keep in mind that too many attributes can lead to problems and that too many choices can be problematic for the respondent (tired, time-consuming, ...).

XLSTAT can generate several unique designs, which is an advantage especially when a large sample of people are interviewed. As the number of different combinations is greater, the analysis designs will be more robust in the analysis of the effects. In addition, including different designs reduces the impact of psychological context and order effects.

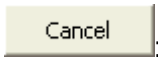
XLSTAT-Conjoint allows obtaining a global table for MaxDiff analysis but also individual tables for each respondent and each comparison in separated Excel sheets. References are also included so that when a respondent select a profile in an individual sheet, it is directly reported in the main table.

Dialog box

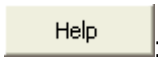
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Analysis name: Enter the name of the analysis you want to perform.

Attributes: Select the attributes that will be tested during this analysis.

Number of respondants: Enter the number of expected individuals who will respond to the MaxDiff analysis.

Number of comparisons (combinations): Enter the maximum number of comparison to be presented to the individual respondents.

Attributes (Choices) per comparison (combination): Enter the number of attributes per comparison.

Terminology: Choose among the proposed alternatives, the terms that best correspond to your case.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selections contains a label.

Outputs tab:

Data format: Choose from one of the two formats in which you want your design to be displayed: Combination/Respondent or Respondent/Combination.

Number of designs: Check this option if you want to generate multiple designs.

- **Respondents by design**: Enter a vector of size equal to the number of designs, containing the number of respondents per design.
- **Comparisons by design**: Enter a vector of size equal to the number of designs, containing the number of combinations per design.

Print individual sheets: Activate this option to print individual sheets for each respondent. Each sheet will include a table for each comparison. The respondent has to enter any value close to the best (on the right) and worst (on the left) attributes. Two assignment options are available; the fixed option displays the comparisons in the same order for all individuals; the random option displays the comparisons in random orders (different from one respondent to another).

Include references: Activate this option to include references between the main sheet and the individual sheets. When an individual enter his chosen code in the individual sheet, the result is automatically displayed in the main sheet of the analysis.

Results

Variable information: This table displays all the information relative to the attributes.

MaxDiff analysis design: This table displays the comparisons presented to the respondent. Each row is associated to a comparison of attributes. Empty cells associated to each individual respondent are also displayed. Respondent have to enter the code associated to the choice made (1 to number of attributes per comparisons). Two columns per respondent have to be filled (best and worst).

Run the analysis: Once you filled in the MaxDiff design with individuals responses, you can click on the "Run the analysis" button to automatically launch the interface and run a MaxDiff analysis.

Individual _Res sheets: When the "Print individual sheets" option is activated, these sheets include the name of the analysis, the individual number and tables associated to the comparisons with the profiles to be compared. Individual respondents should enter the code associated to their choice in the bottom right of each table.

Example

An example of MaxDiff analysis is available on XLSTAT Help Center:

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

References

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.

MaxDiff analysis

Use this tool to run a MaxDiff analysis. This tool is included in the XLSTAT- Conjoint module; it must be applied on design of experiments for MaxDiff analysis generated with XLSTAT-Conjoint.

In this section:

Description

MaxDiff or Maximum Difference Scaling is a method introduced by Jordan Louvière (1991) that allows obtaining importance of attributes. Attributes are presented to a respondent who must choose to best and worst attributes (most important / least important).

This tool allows you to carry out the step of analyzing the results obtained after the collection of responses from a sample of people. This analysis can only be done once the attributes have been defined, the design has been generated, and the respondent responses have been collected.

In the case of MaxDiff models, respondents must choose between selections of attributes. Thus, a number of choices is given to all respondents (we select an attribute from a number of attributes).

Analysis of these choices can be done using a conditional logit model or a hierarchical Bayes algorithm which gives individual results.

Conditional logit model

The conditional logit model is based on a model similar to that of the logistic regression except that instead of having individual characteristics, there will be characteristics of the different alternatives proposed to the respondents.

The probability that respondent i chooses product j is given by:

$$P_{ij} = \frac{e^{\beta^T z_{ij}}}{\sum_k e^{\beta^T z_{ik}}}$$

From this probability, we calculate a likelihood function:

$$l(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \log(P_{ij})$$

With y being a binary variable indicating the choice of respondent i for product j and J being the number of choices available to each respondent.

To estimate the model parameters β (the coefficients of the linear function), it seeks to maximize the likelihood function. Unlike linear regression, an exact analytical solution does not exist. It is therefore necessary to use an iterative algorithm. XLSTAT-Conjoint uses a Newton-Raphson algorithm.

To avoid linear dependency, we arbitrarily set the utility for the first item to zero and estimate the utility of all other items with respect to that first item held constant at zero.

Hierarchical Bayes model

Parameters are estimated at the individual level using an iterative method (Gibbs sampling) taking into account each respondent's choice but also the global distribution of the choices. The obtained individual importance will be more precise.

The MaxDiff analysis allows obtaining individual MaxDiff score for each respondent and each attribute.

The model coefficients are obtained using the HB model with X as input for best choices and $-X$ for worst choices. Then, these coefficients are transformed to obtain MaxDiff scores. They are centered then transformed using that formula: $\frac{\exp(\beta)}{\exp(\beta)+nb_{alter}-1}$ with nb_{alter} being the number of alternatives proposed in each choice task. Then the scores are rescaled in order to sum to 100.

Consistency of participant responses

When participants complete sets of multiple choices, the consistency of their responses can be quantified. If the choices are not deemed consistent, the respondent may then be classified as providing "bad" data and excluded from the analysis. XLSTAT offers two indices for this:

- The RLH (Root LikeliHood). This index only exists in the hierarchical Bayes model. It serves as an indicator of the consistency of choices among respondents. The RLH is a probabilistic expression of the goodness of fit of the data to the fitted model. It can vary between 0 and 1 (0-100%), and higher values indicate greater consistency of choices. It is calculated as follows:

$$RLH = \left(\prod_{i=1}^{q} L_i\right)^{1/q}$$

where L_i is the likelihood of question i given by the model, and q is the number of questions.

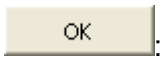
- The ErrVarNorm index. This is a measure of the variance in each participant's choices. For this index, the best respondent is considered the reference for all other participants in the study. The ErrVarNorm index takes values between 0 and 1 and increases with the consistency of the respondent's answers. It is calculated as follows:

$$ErrVarNorm = \frac{\sum_{i=1}^{p} x_i^2}{Max_{j=1}^{m}(\sum_{i=1}^{p} x_i^2)}$$

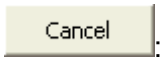
where p is the number of attributes seen and x_i is the score $B - W$ (difference between the number of times each product was chosen as best (Best), and as worst (Worst)) for the attribute i.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Automatically load data  : In order to run a MaxDiff analysis you need to load two data tables: a table with the respondent responses and a table containing the different choices (or combinations of choices). If the MaxDiff design has been generated with XLSTAT , then you can load the two tables automatically. To do that, click on the "magic stick" button and select any cell in the sheet containing the MaxDiff design generated by XLSTAT. In order to correctly load your data, please avoid any manual modification (e.g. adding extra rows or columns) of the sheet containing the XLSTAT MaxDiff design. You can also manually load your data using the fields of the MaxDiff analysis dialog box.

Responses: Select the responses that have been given by respondents. If headers have been selected, please check the option "Variable labels" is enabled. This selection corresponds to the right part of the MaxDiff analysis design table generated with the "design of MaxDiff analysis" tool of XLSTAT- Conjoint.

Choice table: Select the choices that have been presented to the respondents. If headers have been selected, please check the option "Variable labels" is enabled. This selection corresponds to the left part of the MaxDiff analysis design table generated with the "design of MaxDiff analysis" tool of XLSTAT-Conjoint. Do not select the first column of the table.

Terminology: Choose among the proposed alternatives, the terms that best correspond to your study.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selections contains a label.

Response weights: Activate this option if response weights are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection.

Options tab:

Method: Select the method to be used for estimation from Logit or Hierarchical Bayes.

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95%.

Bayesian options: the number of iterations for the burn-in period and the maximal time for the hierarchical Bayes algorithm can be modified.

Stop conditions: the number of iterations and the convergence criterion until convergence of the algorithm can be modified.If the number of iterations is reached and if $Abs(mean(BetaOld - BetaNew))$ (where $BetaOld$ sets the value of the coefficients at iteration $k-1$ and $BetaNew$ set their value to the iteration k) does not reach the convergence value so the algorithm is stopped.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Outputs tab:

Counts analysis: Activate this option to display the counts analysis for all respondents.

- **Individual results**: Activate this option to display the counts analysis for each respondent.

ErrVarNorm index: Activate this option to display a table including the ErrVarNorm index for each respondent.

- **Remove individuals**: Activate this option to remove respondents with an ErrVarNorm index lower than a chosen value (default 0.3) in the rest of the analysis.

In the case of the conditional logit model:

Utilities: Activate this option to display the table allowing you to evaluate the utility associated with each attribute.

In the case of the hierarchical Bayesian model:

MaxDiff Scores: Activate this option to display the importance scores associated with each attribute and for each respondent. Descriptive statistics are also displayed. This importance score is equivalent to the probability score normalized to 100%.

Probability Scores: Activate this option to display the probability scores associated with each attribute and for each respondent.

Model coefficients: Activate this option to display the HB model coefficients.

- **Individual results**: Activate this option to display the coefficients of the HB model for each respondent.

Goodness of fit coefficients: Activate this option to display a series of statistics for the independent model (corresponding to the case where the linear combination of explanatory variables is reduced to a constant) and for the adjusted model.

RLH per respondent: Activate this option to display the RLH per respondent.

Results

Variable information: This table displays all the information relative to the used attributes.

Counts analysis: These tables summarize the results of the MaxDiff survey by showing how many times each attribute has been chosen as best and worst, first for the entire population and then for each respondent. The third column of these tables correspond to the difference of the best and worst frequencies. Finally, the last column of the table is the results of the square root of the number of times each product was chosen as best divided by the number of times each product was chosen as worst, which gives a first idea of the results of the analysis.

$$\sqrt{Most/Least}$$

ErrVarNorm Index: This table presents the ErrVarNorm index for each respondent.

Utilities: If the chosen model is "logit", this table is displayed to make it possible to evaluate the utility associated with each attribute.

The following results are displayed in the case of the Bayesian model Hierarchical.

MaxDiff Scores: This table displays the importance scores associated with each attribute and for each respondent. Descriptive statistics tables for all respondents are also available.

Probability scores: This table displays the probability scores associated with each attribute and for each respondent.

Model coefficients: This table displays the HB model coefficients.

Goodness of fit coefficients: This table displays a series of statistics for the independent model (corresponding to the case where the linear combination of explanatory variables

reduces to a constant) and for the adjusted model.

- **Observations**: The total number of observations taken into account (sum of the weights of the observations);
- **Sum of weights**: The total number of observations taken into account (sum of the weights of the observations multiplied by the weights in the regression);
- **-2 Log(Like.)** : The logarithm of the likelihood function associated with the model;
- **rlh**: root likelihood. This value varies between 0 and 1, the value of 1 being a perfect fit.
- **rlh by respondent**: The RLH (Root LikeliHood) value is an index from 0 to 1. The higher the RLH value of a respondent, the more consistently the respondent answered the choice questions."

respondent results are then displayed.

Example

An example of MaxDiff analysis is available on XLSTAT Help Center:

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

References

Jaeger, S.R., Llobell, F. (Eurosense 2024). Best-worst scaling (BWS) in food consumer science: Why, when and how?

Llobell, F., Choisy, P., Chheang, S.L., Jaeger, S.R. (June 5, 2024). Best Worst Scaling in sensory analysis: how to detect atypical respondents?. Sensometrics, Paris, France.

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.

Monotone regression (MONANOVA)

Use this tool to apply a monotone regression or MONANOVA model. Advanced options let you choose the constraints on the model and take into account interactions between factors. This tool is included in the module XLSTAT- Conjoint.

In this section:

Description

The MONANOVA model is part of the XLSTAT-Conjoint module.

Monotone regression and the MONANOVA model differ only in the fact that the explanatory variables are either quantitative or qualitative. These methods are based on iterative algorithms based on the ALS (alternating least squares) algorithm. Their principle is simple, it consists of alternating between a conventional estimation using linear regression or ANOVA and a monotonic transformation of the dependent variables (after searching for optimal scaling transformations).

The MONANOVA algorithm was introduced by Kruskal (1965) and the monotone regression and the works on the ALS algorithm are due to Young *& al.* (1976).

These methods are commonly used as part of the full profile conjoint analysis. XLSTAT-Conjoint allows applying them within a conjoint analysis (see chapter on conjoint analysis based on full profiles) as well as independently.

The monotone regression tool (MONANOVA) combines a monotonic transformation of the responses to a linear regression as a way to improve the linear regression results. It is well suited to ordinal dependent variables.

XLSTAT-Conjoint allows you to add interactions and to vary the constraints on the variables.

Method

Monotone regression combines two stages: an ordinary linear regression between the explanatory variables and the response variable and a transformation step of the response variables to maximize the quality of prediction.

The algorithm is:

1. Run an OLS regression between the response variable Y and the explanatory variables X. We obtain the β coefficients.
2. Calculation of the predicted values of Y: $Pred(Y) = \beta * X$
3. Transformation of Y using a monotonic transformation (Kruskal, 1965) so that $Pred(Y)$ and Y are close (using optimal scaling methods).
4. Run an OLS Regression between Y_trans and the explanatory variables X. This gives new values for the β.
5. Steps 2 through 4 are repeated until the change in R^2 from one stage to another is smaller than the convergence criterion.

Goodness of fit (MONANOVA)

In the context of MONANOVA, additional results are available. These results are generally associated with a multivariate analysis but as we are in the case of a transformation of the responses, their presence is necessary. Instead of using the squared canonical correlations between measures, we use the R^2. XLSTAT-Conjoint calculates the Wilks' lambda, Pillai's trace, the trace of Hotelling-Lawlet and Roy largest root using a matrix with largest eigenvalue equal to the R^2 and 0 for other eigenvalues. The largest root of Roy gives a lower bound for the p-value of the model. Other statistics are upper bounds on the p-value of the model.

Interactions

By interaction is meant an artificial factor (not measured) which reflects the interaction between at least two measured factors. For example, if we carry out treatment on a plant, and tests are carried out under two different light intensities, we will be able to include in the model an interaction factor treatment*light which will be used to identify a possible interaction between the two factors. If there is an interaction between the two factors, we will observe a significantly larger effect on the plants when the light is strong and the treatment is of type 2 while the effect is average for weak light, treatment 2 and strong light, treatment 1 combinations.

To make a parallel with linear regression, the interactions are equivalent to the products between the continuous explanatory values although here obtaining interactions requires nothing more than simple multiplication between two variables. However, the notation used to represent the interaction between factor A and factor B is $A * B$.

The interactions to be used in the model can be easily defined in XLSTAT- Conjoint.

Constraints for qualitative predictors

During the calculations, each factor is broken down into a sub-matrix containing as many columns as there are categories in the factor. Typically, this is a full disjunctive table. Nevertheless, the breakdown poses a problem: if there are g categories, the rank of this sub-matrix is not g but $g-1$. This leads to the requirement to delete one of the columns of the sub-matrix and possibly to transform the other columns. Several strategies are available depending on the interpretation we want to make afterwards:

1) **a1=0**: the parameter for the first category is null. This choice allows us force the effect of the first category as a standard.

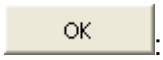
2) **an=0**: the parameter for the last category is null. This choice allows us force the effect of the last category as a standard.

3) **Sum (ai) = 0**: the sum of the parameters is null. This choice forces the constant of the model to be equal to the mean of the dependent variable when the design is balanced.

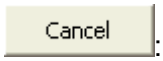
Note: even if the choice of constraint influences the values of the parameters, it has no effect on the predicted values and on the different fitting statistics.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Dependent variables:

Quantitative: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

X / Explanatory variables:

Quantitative: Select the quantitative explanatory variables in the Excel worksheet. The data selected must be of type numeric. If the variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Select the qualitative explanatory variables (the factors) in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2,…).

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

Options tab:

Fixed constant: Activate this option to fix the constant of the regression model to a value you then enter (0 by default).

Tolerance: Activate this option to prevent the OLS regression calculation algorithm taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95%.

Constraints: Details on the various options are available in the description section.

a1 = 0: Choose this option so that the parameter of the first category of each factor is set to 0.

an = 0: Choose this option so that the parameter of the last category of each factor is set to 0.

Sum (ai) = 0: for each factor, the sum of the parameters associated with the various categories is set to 0.

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the ALS algorithm. The calculations are stopped when the maximum number if iterations has been exceeded. Default value: 100.
- **Convergence**: Enter the maximum value of the evolution of R^2 from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

- **Check for each Y separately**: Choose this option to remove the observations with missing data in the select Y (dependent) variables, only if the Y of interest has a missing data.
- **Across all Ys**: Choose this option to remove the observations with missing data in the Y (dependent) variables, even if the Y of interest has no missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Observation details: Activate this option to display detailed outputs for each respondent.

Correlations: Activate this option to display the correlation matrix for quantitative variables (dependent or explanatory).

Analysis of variance: Activate this option to display the analysis of variance table.

Type I/II/III SS: Activate this option to display the Type I, Type II, and Type III sum of squares tables. Type II table is only displayed if it is different from Type III.

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

Charts tab:

Regression charts: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.

Transformation plot: Activate this option to display the monotone transformation of the response plot.

Results

XLSTAT displays a large number tables and charts to help in analyzing and interpreting the results.

Summary statistics: This table displays descriptive statistics for all the variables selected. For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

Correlation matrix: This table displays the correlations between the explanatory variables.

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.

- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).
- R^2: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y}_i)^2}, \text{ where } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i$$

The R^2 is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer R^2 is to 1, the better is the model. The problem with the R^2 is that it does not take into account the number of variables used to fit the model.

- **Adjusted** R^2: The adjusted determination coefficient for the model. The adjusted R^2 can be negative if the R^2 is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

$$\hat{R}^2 = 1 - (1 - R^2) \frac{W - 1}{W - p - 1}$$

The adjusted R^2 is a correction to the R^2 which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.
- **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows::

$$MAPE = \frac{100}{W} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- **DW**: The Durbin-Watson statistic is defined by:

$$DW = \frac{\sum_{i=2}^{n} [(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}$$

This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- Cp: Mallows Cp coefficient is defined by:

$$Cp = \frac{SCE}{\hat{\sigma}} + 2p^* - W$$

where SSE is the sum of the squares of the errors for the model with p explanatory variables and $\hat{\sigma}$ is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to $p*$, the less the model is biased.

- **AIC**: Akaike's Information Criterion is defined by:

$$AIC = W \ln\left(\frac{SCE}{W}\right) + 2p^*$$

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz's Bayesian Criterion is defined by:

$$SBC = W \ln\left(\frac{SCE}{W}\right) + \ln(W)p^*$$

This criterion, proposed by Schwarz (1978) is similar to the AIC and, like this, the aim is to minimize it.

- **PC**: Amemiya's Prediction Criterion is defined by:

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}$$

This criterion, proposed by Amemiya (1980) is used, like the adjusted R^2 to take account of the parsimony of the model.

- **Press RMSE**: Press' statistic is only displayed if the corresponding option has been activated in the dialog box. It is defined by:

$$Press = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i(-i)})^2$$

where $\hat{y}_{i(-i)}$is the prediction for observation i when the latter is not used for estimating parameters. We then get:

$$\text{Press RMCE } = \sqrt{\frac{\text{Press}}{W - p^*}}$$

Press's RMSE can then be compared to the RMSE. A large difference between the two shows that the model is sensitive to the presence or absence of certain observations in the model.

- **Iteration**: Number of iteration until convergence of the ALS algorithm.

Goodness of fit coefficients (MONANOVA): In this table are shown the statistics for the fit of the regression model specific to the case of MONANOVA. These statistics are the Wilks' lambda, the Pillai's trace, the trace of Hotelling-Lawlet and the largest root of Roy. For more details on these statistics, see the description part of this help.

If the Type I/II/III SS (SS: Sum of Squares) option is activated, the corresponding tables are displayed.

The table of **Type I SS** values is used to visualize the influence that progressively adding explanatory variables has on the fitting of the model, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. The sums of squares in the Type I table always add up to the model SS. Note: the order in which the variables are selected in the model influences the values obtained.

The table of **Type II SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. Type II SS are not recommended in unbalanced designs but we display them as some users might need them. It is identical to Type III for balanced designs.

The table of **Type III SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. While Type II SS depends on the number of observations per cell (cell means combination of categories of the factors), Type III does not and is therefore preferred.

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

The **parameters of the model** table displays the estimate of the parameters, the corresponding standard error, the Student’s t, the corresponding probability, as well as the confidence interval

The table of **standardized coefficients** (also called β coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the observed value of the dependent variable, the transformed value of the dependant variable, the model's prediction, the residuals, and the confidence intervals. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the random values being larger.

The **charts** which follow show the transformation of the dependant variable.

Example

An example of MONANOVA is available on XLSTAT Help Center:

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

References

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.

Conditional logit model

Use conditional logit model to model a binary variable using quantitative and/or qualitative explanatory.

In this section:

Description

The conditional logit model is part of the XLSTAT-Conjoint module.

The conditional logit model is based on a model similar to that of the logistic regression. The difference is that all individuals are subjected to different situations before expressing their choice (modeled using a binary variable which is the dependent variable). The fact that the same individuals are used in taken in account by the conditional logit model (NB: the observations are not independent within a block corresponding to same individual).

The conditional logit model is a method mostly used in conjoint analysis, it is nevertheless useful when analyzing a certain type of data. It is McFadden (1973) who introduced this model. Instead of having one line per individual like in the classical logit model, there will be one row for each category of the variable of interest. If one seeks to study transportations, for example, there will be four types of transports (car / train / plane / bike), each type of transport have characteristics (their price, their environmental costs...) but an individual can choose only one of four transportations. As part of a conditional logit model, all four options are presented to each individual and the individual choose his preferred option. We have for N individuals, $N * 4$ rows with 4 rows for each individual associated with each transportation. The binary response variable will indicate the choice of the individual (1) and 0 if the individual did not choose this option.

In XLSTAT-Conjoint, you will also have to select a column associated with the name of the individuals (with 4 lines per individual in our example). The explanatory variables will also have $N * 4$ lines.

Method

The conditional logit model is based on a model similar to that of the logistic regression except that instead of having individual characteristics, there will be characteristics of the different

alternatives proposed to the individuals.

The probability that individual i chooses product j is given by:

$$P_{ij} = \frac{e^{\beta^T z_{ij}}}{\sum_k e^{\beta^T z_{ik}}}$$

From this probability, we calculate a likelihood function:

$$l(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \log(P_{ij})$$

With y being a binary variable indicating the choice of individual i for product j and J being the number of choices available to each individual.

To estimate the model parameters β (the coefficients of the linear function), it seeks to maximize the likelihood function. Unlike linear regression, an exact analytical solution does not exist. It is therefore necessary to use an iterative algorithm. XLSTAT-Conjoint uses a Newton-Raphson algorithm.

Goodness of fit (conditional logit)

Some specific goodness of fit indexes are displayed for the conditional logit model.

- Likelihood ratio R: $R = -2(\log(L) - \log(L_0))$

- Upper bound of the likelihood ratio 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: $Estrella = 1 - (1 - \frac{R}{U})^{\frac{U}{N}}$

- Adjusted Estrella: $Adj.Estrella = 1 - \left(\frac{\log(L)-k}{\log(L_0)}\right)^{\frac{2}{N}\log(L_0)}$

- Veall-Zimmermann: $VZ = \frac{R(U+N)}{U(R+N)}$

With N being the sample size and K being the number of predictors.

Constraints for qualitative predictors

During the calculations, when qualitative predictors are selected, each factor is broken down into a sub-matrix containing as many columns as there are categories in the factor. Typically, this is a full disjunctive table. Nevertheless, the breakdown poses a problem: if there are g categories, the rank of this sub-matrix is not g but $g-1$. This leads to the requirement to delete one of the columns of the sub-matrix and possibly to transform the other columns. Several strategies are available depending on the interpretation we want to make afterwards:

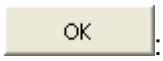
1) **a1=0**: the parameter for the first category is null. This choice allows us force the effect of the first category as a standard.

2) **Sum (ai) = 0**: the sum of the parameters is null.

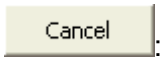
Note: even if the choice of constraint influences the values of the parameters, it has no effect on the predicted values and on the different fitting statistics.

Dialog box

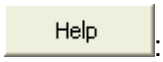
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Response variable: Select the response variable you want to model. If headers have been selected, please check the option "Variable labels" is enabled. This variable has to be a binary variable.

Subject variable: Select the subject variable corresponding to the name of the individuals. If headers have been selected, please check the option "Variable labels" is enabled.

Explanatory variables:

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If the variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selections (data, other group) contains a label.

Observation weights: Activate this option if observations weights are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection.

Options tab:

Tolerance: Enter the value of the tolerance threshold below which a variable will automatically be ignored.

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95%.

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the Newton-Raphson algorithm. The calculations are stopped when the maximum number if iterations has been exceeded. Default value: 100.
- **Convergence**: Enter the maximum value of the evolution of the log of the likelihood from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the explanatory variables correlation matrix.

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Type III analysis: Activate this option to display the type III analysis of variance table.

Standardized coefficients: Activate this option if you want the standardized coefficients (β coefficients) for the model to be displayed.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

Charts tab:

Regression charts: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.

Results

XLSTAT displays a large number tables and charts to help in analyzing and interpreting the results.

Summary statistics: This table displays descriptive statistics for all the variables selected. For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

Correlation matrix: This table displays the correlations between the explanatory variables.

Goodness of fit coefficients: This table displays a series of statistics for the independent model (corresponding to the case where the linear combination of explanatory variables reduces to a constant) and for the adjusted model.

- **Observations**: The total number of observations taken into account (sum of the weights of the observations);
- **Sum of weights**: The total number of observations taken into account (sum of the weights of the observations multiplied by the weights in the regression);
- **DF**: Degrees of freedom;
- **-2 Log(Like.)**: The logarithm of the likelihood function associated with the model;
- R^2**(McFadden)**: Coefficient, like the R^2, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model;
- R^2**(Cox and Snell)**: Coefficient, like the R^2, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model raised to the power $\frac{2}{S_w}$, where S_w is the sum of weights.
- R^2**(Nagelkerke)**: Coefficient, like the R^2, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to ratio of the R^2 of Cox and Snell, divided by 1 minus the likelihood of the independent model raised to the power $\frac{2}{S_w}$;
- **AIC**: Akaike's Information Criterion;
- **SBC**: Schwarz's Bayesian Criterion.
- **Iteration**: Number of iteration to reach convergence.

Goodness of fit indexes (conditional logit): In this table are shown the goodness of fit statistics specific to the case of the conditional logit model. For more details on these statistics, see the description part of this help.

Test of the null hypothesis $H_0 : Y = p_0$: The H_0 hypothesis corresponds to the independent model which gives probability p_0 whatever the values of the explanatory variables. We seek to check if the adjusted model is significantly more powerful than this model. Three tests are available: the likelihood ratio test (-2 Log(Like.)), the Score test and the Wald test. The three statistics follow a χ^2 distribution whose degrees of freedom are shown.

Type III analysis: This table is only useful if there is more than one explanatory variable. Here, the adjusted model is tested against a test model where the variable in the row of the table in question has been removed. If the probability $Pr > LR$ is less than a significance threshold which has been set (typically 0.05), then the contribution of the variable to the adjustment of the model is significant. Otherwise, it can be removed from the model.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can easily be seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the observed value of the dependent variable, the model's prediction, the same values divided by the weights, the standardized residuals and a confidence interval.

Example

An example of conditional logit model is available on XLSTAT Help Center:

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

References

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

Feature extraction

Feature extraction is used to reduce the number of resources required to describe a large set of textual data. It is a general term describing methods of constructing variable combinations in order to avoid these problems while still describing the data with sufficient accuracy. The “extracted features” are commonly used in document classification methods where the occurrence frequency of each word in a document is used as a feature for training a classifier.

In this section:

Description

Since machine learning algorithms operate on a numeric feature space, the expected input is a two-dimensional array where rows are observations (documents) to classify and columns are features (words). Consequently, to perform machine learning on textual data, we need to transform our documents into vector representations such that we can apply numeric machine learning. The process of encoding documents in a numeric feature space is called *feature extraction* or more simply, vectorization and is an essential first step in Natural Language Processing (NLP)

One simple encoding for the vectorizing process is the “bag-of-words” model, better known as vector space model. In this model, a text (such as a sentence or a document) is represented as a multiset of its words, disregarding grammar and even word order. A classical output with this scheme is the document-term matrix.

Document-term matrix

Document-term matrix (**DTM**) uses all the tokens in the dataset as vocabulary. It is a mathematical matrix object that describes the frequency of terms that occur in a collection of documents. In a document-term frequency matrix, each row in the matrix corresponds to a document and each column corresponds to a term (word) in the document. Each cell represents the frequency (number of occurrences) of the corresponding word in the corresponding document. Prior to that, a tokenization step is performed to separate words from the document by using white space as the delimiter. Moreover, many different filtering combinations are applied to remove words that do not have any significant importance in building the matrix. This

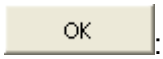
process is called stop words removal (stop words are words like: a, the, that and so on). Other filtering procedures can be performed such as removing sparse terms (terms that are not present above a certain proportion over the whole documents), or stemming to remove inflexional endings from the necessary words (thus reducing inflected words to their word stem).

Term-document matrix and word cloud

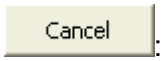
In a term-document frequency matrix (**TDM**), each row in the matrix corresponds to a term and each column corresponds to a document. Each cell represents the frequency (number of occurrences) of the corresponding word in the corresponding document.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Worksheet: Select a table made up of N documents (one document stored in each cell) from an Excel spreadsheet. If the "document labels" option has been selected, check that cells containing labels have been selected.

Document files (.txt): Select multiple text files (WINDOWS version) or a folder containing them (MAC version). Each file is read as a single document.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first column of the data selections contains document labels.

Document labels: Activate this option if document labels are available. Then select the corresponding data. If the "Column labels" option is activated, you need to include a header in the selection. If this option is not activated, the observation labels are automatically generated by XLSTAT (Doc1, Doc2, etc.).

Options tab:

Preprocessing sub-tab:

Vocabulary filtering

Stop words list: Activate this option to exclude a list of stop words (insignificant words in a text) contained in documents (English language by default).

Remove punctuation: Activate this option to delete punctuation marks in documents.

Remove numbers: Activate this option to remove numbers in documents.

Text normalization

Stemming: Activate this option to reduce words to their common stem (English stemmer by default).

Intermediate form sub-tab:

Term filtering

Remove sparse terms: Activate this option to delete terms whose proportion of presence is lower than 100 x (1- *value*) % over the whole documents (*value* is 0.95 by default).

Minimum frequency: Activate this option to skip terms occurring less than *value* time(s) over the whole documents (*value* is 2 by default)

Maximum frequency: Activate this option to skip terms occurring more than *value* time(s) over the whole documents (*value* shall be greater or equal to the *minimum frequency* parameter if activated)

Maximum number: Activate this option to set a maximum number of words to be included in the DTM or TDM. The least frequent terms will be dropped.

Outputs tab:

Term-document matrix: Activate this option to display the TDM in the XLSTAT result sheet.

Document-term matrix: Activate this option to display the DTM in the XLSTAT result sheet.

Export document-term matrix (DTM) or term-document matrix (TDM): Activate this option to specify a folder path where to export the document-term matrix or term-document matrix as comma-separated values format (CSV).

Charts tab:

Word cloud: Activate this option to display the word cloud representing all documents.

Results

The DTM or the TDM is displayed. The exported DTM or TDM (if the corresponding option is chosen) has no limitation for the maximum number of terms allowed to be contained (useful when exceeding the Excel limitation in the result sheet).

At the end of the term-document matrix, the following button is displayed:  . Click on this button to automatically open the pre-filled dialog box of (Word cloud) to create and personalize word cloud(s).

Example

An example based on data collected from the Internet Movie DataBase (IMBD) is permanently available on the XLSTAT Help Center. To download this data, go to:

https://help.xlstat.com/customer/en/portal/articles/2937383-feature-extraction-tutorial-in-excel?b_id=9283

References

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.

Latent Semantic Analysis (LSA)

Use Latent Semantic Analysis (LSA) to discover the hidden and underlying (latent) semantics of words in a corpus of documents.

In this section:

Description

Latent Semantic Analysis (LSA) allows you to discover the hidden and underlying (latent) semantics of words in a corpus of documents by constructing "concepts" (or "topic") related to documents and terms. The LSA uses an input document-term matrix that describes the occurrence of group of terms in documents. It is a sparse matrix whose lines correspond to "documents" and whose columns correspond to "terms".

Uses of LSA

There are several applications for LSA, including:

- Compare documents in the low-dimensional space (data clustering, document classification).
- Find similar documents across languages, after analyzing a base set of translated documents (cross language retrieval).
- Find relations between terms (synonymy and polysemy).

Principle of LSA

Latent Semantic Analysis (LSA) relies on a document-term matrix. The elements of this matrix contain the occurrences of the different terms in each document.

This matrix is then used to make associations between documents and concepts (from the terms), and thus to link the documents to one another semantically. To do this, we perform different mathematical operations on the matrix, in the following order:

- Calculation of the singular value decomposition (SVD) of the M document-term matrix to reveal the eigen vector spaces of the documents and terms to get the following relation $M = D \times S \times T^T$
- Selection of the k first singular values (in order of importance) of the S diagonal matrix and the corresponding columns in the T and D matrices. The D matrix represents the document vectors in the terms space and the T matrix the term vectors in the documents space. Hereunder is the expression of the new truncated SVD:

$$M_k = D_k \times S_k \times T_k^T$$

Similar terms can then be found by calculating the cosine similarity between two columns of the M_k reduced rank matrix which is strictly equivalent to the cosine similarity between the corresponding columns of $S_k \times T_k^T$.

The same principle applies to finding similar documents by calculating the cosine similarity between two rows of the M_k reduced rank matrix which is strictly equivalent to the cosine similarity between the corresponding lines of $D_k \times S_k$.

$$sim_{cosinus}(a_i, a_j) = \frac{a_i a_j^T}{||a_i||||a_j||}$$

with $\{a_i, a_j\}$ possibly being a pair of terms or documents.

Interpreting the results

Representation of the terms in the semantic k space makes it possible to visually interpret the similarities between the terms on the one hand, and between the documents on the other hand.

Indeed, whether it is the representation of documents or terms in the latent semantic space, two vectors very far in the space of the original matrix may appear close in the vector space reduced to k latent dimensions because the rank reduction has the effect of merging dimensions associated with terms / documents with similar meaning.

Number of factors

Two methods are commonly used for determining the number of factors to be used for interpreting the results:

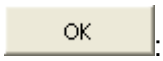
The *scree test* (Cattell, 1966) is based on the decreasing curve of eigenvalues. The number of factors to be kept corresponds to the first turning point found on the curve.

We can also use the cumulative variability percentage represented by the factor axes and decide to use only a certain percentage.

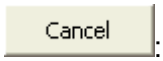
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various

elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.


: Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Document-term matrix: select a table including N documents described by P terms. If term labels have been selected, please check that the option "Term labels" is activated.

Documents weights: enable this option if you want to weight the documents. If you do not activate this option, the weights will all be considered 1. These must imperatively be greater than or equal to 0. If a column header has been selected, check that the "Term labels" option is activated.

Range: activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: activate this option to display the results in a new worksheet of the active workbook.

Workbook: activate this option to display the results in a new workbook.

Term labels: activate this option if the first row of the data selections (Document-term matrix, Document labels, Document weights) includes a header.

Document labels: activate this option if observations labels are available. Then select the corresponding data. If the "Term labels" option is activated you need to include a header in the

selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Doc1, Doc2 …).

Options tab:

Number of topics: enter the number of topics considered for which Latent Semantic Analysis will be applied.

Document clustering: enable this option if you want to create document classes in the created semantic space. These classes can be displayed via the **Color by class** option located below the **Document-document correlation matrix** check box in the charts tab.

Term clustering: enable this option if you want to create term classes in the created semantic space. These classes can be displayed via the **Color by class** option located below the **Term-term correlation matrix** check box in the charts tab.

Type of clustering: you can activate one of the two options to select the type of clustering related the two clustering options above.

- **Hard**: choose this option to perform a classification in the new created semantic space in which each element (term / document) can belong to only one topic at a time to represent a class (hard clustering).
- **Fuzzy**: choose this option to perform a classification in the new semantic space created in which each element (term / document) can belong to several topics at once to represent a class (Soft clustering).

Stop conditions:

- **Iterations**: enter the maximum number of iterations for the SVD algorithm. Calculations are stopped as soon as the maximum number of iterations is exceeded. Default value: 500. If the stop condition is not selected, the algorithm will iterate to the maximum rank of the input document-term matrix.

Data options tab:

Missing data:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove the observations: Activate this option to remove observations with missing data.

Replace missing data by 0: Activate this option to replace missing data by 0.

Outputs tab:

Summary table: enable this option to display the summary of Latent Semantic Analysis. This includes an enumeration of the number of terms and documents for each topic as well as the table (*scree plot*) of the eigenvalues related to the latent topics resulting from the

decomposition. The values are displayed in descending order of amplitude and variability explained.

Topics table: enable this option to display the table of terms that compose each topic.

- **Max. terms/topic**: enable this option to specify the maximum number of terms to display in the topic table. This value will also be applied to the display of correlation matrices graphs.

Nearest neighbor terms: enable this option to display the nearest neighbor terms table to a given term in the created semantic space.

- **Number of terms**: enable this option to specify the number of terms on which to calculate the nearest neighbor terms table.
- **Number of nearest terms**: enable this option to specify the number of nearest neighbors to display in the nearest neighbor terms table. These will be displayed from left to right in descending order of similarity.

Charts tab:

Scree plot: enable this option to display the graph (*scree plot*) of the eigenvalues related to the latent topics resulting from the decomposition. The values are displayed in descending order of amplitude and variability explained.

Term-Term correlation matrix: enable this option to display the correlation matrix representing the term-term correlations (similarities) in the new semantic space.

- **Color by class**: enable this option to color terms related to each class (topic) from the previous classification.

Document-Document correlation matrix: enable this option to display the correlation matrix representing the document-document correlations (similarities) in the new semantic space.

- **Color by class**: enable this option to color documents related to each class (topic) from the previous classification.

Legend: enable this option so that the legend of the different correlation matrices is displayed on the graphs. The latter will not be available when the *Color by class* option is active.

Results

Summary table: the summary table shows the total number of document-terms composing them for each topic. The user has the opportunity thereafter to display all of these in the graphs related to *the correlation matrices* as well as in the *topic table*.

The eigenvalues and the corresponding *scree plot* are also displayed. The cumulative variance provides an indication of the relevance of the calculated topics. The higher the latter, the better the approximation resulting from the "truncated" SVD.

Topics table: this table displays the list of terms / topic from left to right in descending order of relationship with the topic concerned.

Nearest neighbor terms: this table displays the *n* nearest neighbors terms related to the term selected in the drop-down list, in descending order of similarity.

Correlation matrices: the correlation graphs (term-term, document-document, term-document) make it possible to visualize the degree of similarity (cosine similarity) between the terms (**Term-term correlation matrix**) or the documents (**Document-document correlation matrix**) or between the terms and documents (**Term-document correlation matrix**) in their respective spaces. The similarities are between 0 and 1, the value 1 corresponding to a perfect similarity in both directions (positive and negative).

Example

A tutorial on how to use Latent Semantic Analysis is available on the XLSTAT Help Center:

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

References

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.

Sentiment analysis

Use this tool to determine the opinion of an English document thanks to the Syuzhet package.

In this section:

Description

Sentiment analysis is the process of extracting an author's emotional intent from the text (Ted Kwarler, 2017). Sentiment analysis allows you to label a comment, a book, or in general a document. The document can be labeled as a positive, negative, or neutral opinion.

When to use sentiment analysis?

Sentiment analysis helps companies to understand customers' reviews or feedback, product review, or analyze comments on the web (as tweets, or posts), and political discussions. In general, sentiment analysis answers "How do people feel about something?".

What does sentiment analysis use?

Sentiment analysis uses a dictionary where terms are scored or categorized in a polarity way (positive, negative, or neutral). Dictionaries use different scales which is why XLSTAT suggests four sentiment dictionaries to assign sentiment values to terms:

- Sentiment analysis with **Bing** dictionary: 6789 English terms are labeled as "negative", "neutral" or "positive" in the Bing dictionary. A term labeled as "negative" get a score of -1, a term labeled as "neutral" get a score of 0, and on the contrary, a term labeled as "positive" get a score of 1.
- Sentiment analysis with **Syuzhet** dictionary: 10748 English terms are rated between -1 and 1 in the Syuzhet dictionary. A term is labeled as "negative" if its score is lower than 0, and on the contrary, a term is labeled as "positive" if its score is greater than 0.
- Sentiment analysis with **AFINN** dictionary: 3382 English terms are rated between -5 and 5 (integer only) in the AFINN dictionary. A term is labeled as "negative" if its score is lower than 0, and on the contrary, a term is labeled as "positive" if its score is greater than 0.
- Sentiment analysis with **NRC** dictionary (emotion scale): This dictionary labels 13901 English terms with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive).

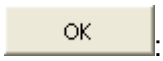
Besides a sentiment dictionary, sentiment analysis needs tokenized documents. XLSTAT suggests using the Feature extraction tool, before going on sentiment analysis to get the document-term matrix.

How is the document score computed?

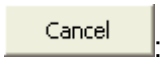
The score of each term present in the document is multiplied by its frequency, then scores are summed to compute the document score.

Dialog box

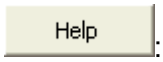
The dialog box is divided into several tabs that correspond to a variety of options, ranging from data selection to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Term frequencies: Select in this field the term frequency matrix. One column corresponds to the frequencies of one term in each document. If the "Column labels" option is activated, you need to include a header in the selection.

Sentiment dictionary: Choose among four sentiment dictionaries (see description).

Custom scores: Select in this field two columns including the term and its score. If you choose the Bing dictionary as the sentiment dictionary, you must enter "negative", "neutral" or "positive". This option allows you to define the sentiment of a term independently of the dictionary previously selected. If the "Column labels" option is activated, you need to include a header in

the selection. For this field, missing values are read as "neutral" or zero. Note: Not available for the NRC dictionary.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet in the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selection contains a label.

Document labels: Activate this option if the document labels are available. Then select the corresponding data. If the "Column labels" option is activated, you need to include a header in the selection. If this option is not activated, the document labels are automatically generated by XLSTAT (Doc1, Doc2, etc.).

Missing data tab:

Remove observations: Activate this option to remove the rows with missing data.

Ignore missing data: Activate this option to ignore missing data.

Outputs tab:

Term frequencies and scores: Activate this option to display a table showing the total frequency and the score of each term included in the term frequency selection. Note: Not available for the NRC dictionary.

Term frequencies and associated emotion: Activate this option to display a table showing the total frequency and the associated emotion of each term including in the term frequencies selection. Note: Only available for the NRC dictionary.

- **Display sentiment terms only**: Activate this option to display only the sentiment terms. Terms with a neutral sentiment, which means their score is equal to zero or they are not associated with an emotion, are not displayed.

Overall emotion frequencies: Activate this option to display the total frequency of each emotion present in all documents. Note: Only available for the NRC dictionary.

Document scores: Activate this option to display a table showing the score of each document (row) according to the sentiment dictionary chosen in the General tab. Note: Only available for the NRC dictionary.

- **Sort by score (descending)**: Activate this option to sort the document scores in descending order.

Emotion frequencies by document: Activate this option to display a table that indicates the frequency of each emotion in each document. Note: Only available for the NRC dictionary.

Result interpretations: Activate this option to display, under the result tables short interpretation.

Charts tab:

Term frequencies: Activate this option to display a bar chart showing the total term frequencies.

- **Minimum frequency**: Enter the minimum frequency a term should have to be displayed in the term frequencies bar chart. We suggest increasing the minimum frequency when the number of terms increases.

Term scores: Activate this option to display a bar chart showing the term score.

Document scores: Activate this option to display a bar chart showing the document score. If the **Sort by score (descending)** option is activated the bar chat is also sorted.

Document scores distribution: Activate this option to display a histogram showing the distribution of the document scores.

Overall emotion frequencies: Activate this option to display a bar chart showing the total emotion frequencies. Note: Only available for the NRC dictionary.

Sentiment-based word cloud: Activate this option to display a word cloud where terms are colored according to their sentiment (positive, negative, or the associated emotion).

- **Maximum terms**: Enter the number maximum of terms to include in the sentiment-based word cloud.

Result interpretations: Activate this option to display, under the charts short interpretation.

Results

Results regarding the document scores: The table and the chart associated with the document scores are displayed to give you a view of the sentiment of each document according to the sentiment dictionary scale. If the **Sort by score (descending)** option is not activated, you can see the evolution of the document scores, especially if the documents are entered chronologically.

Results regarding the document and the associated emotion: With the emotion scale (NRC), a table is displayed showing the frequencies of each emotion present in a document. This table can be completed with the document scores obtained by another sentiment dictionary and allows you to put natural words on the sentiment and intensity of an opinion present in a document.

Result regarding the document scores distribution: The histogram displayed helps to know the frequency of the scores. In case the scores are centered at 0, it means that on average the documents have many neutral words in them. In another hand, if the scores are centered at a value higher (resp. lower) than 0, it means that on average each document has at least a single positive (resp. negative) word in it.

Results regarding the term frequencies: The table and the chart associated with the term frequencies are displayed to give you a view of the total frequency of a term, in other words, it shows the number of occurrences of the term among all documents.

Results regarding the term scores: The table and the chart associated with the term scores are displayed to give you a view of the sentiment of each term according to the sentiment

dictionary scale. In the case of the emotion scale, a term can be associated with zero, one, or several emotions. Neutral terms have a blank case in the "Score" column. Custom scores are shown in bold.

Example

A tutorial on sentiment analysis is available at the XLSTAT Help Center:

https://www.xlstat.com/demo/stm_en

References

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*.

Terms selection

Use this method to perform a regression on a document-term matrix.

In this section:

Description

Terms selection uses the well-known Elastic net regression method and its logistic version. Indeed, it allows you to model quantitative variables but also binomial (typically binary) variables and multinomial variables (qualitative variables with more than two categories).

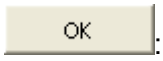
Terms selection is a method used only in the case of text mining, where the document-term matrix replaces the quantitative explanatory variables, and the sentiment vector is the response variable giving the sentiment ("positive", "negative", etc.) of each document or its rate (quantitative indication of the opinion).

The Elastic net regression is based on two fundamental parameters: the mixing parameter α (which is between 0 and 1) and the λ regularization parameter > 0. XLSTAT offers to its users to find the optimal λ parameters by cross-validation.

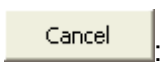
If you want to know more about the Elastic net Regression, see its description. For the Elastic net logistic regression method, please complete the description of the Elastic net Regression with the Logistic regression.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options, ranging from data selection to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Response variable: Select the response variable you want to model. If a column header has been selected, check that the "Column labels" option has been activated.

Response type: Select your type of response variable:

- **Gaussian**: If your response variable is numerical, choose this type to fit a regression model.
- **Poisson**: If your response variable is numerical, choose this type to fit a regression model.
- **Binomial**: If your response variable is binary, choose this type to fit a regression logistic model.
- **Multinomial**: If your response variable includes more than two categories, choose this type to fit a regression logistic model.

Term frequencies: Select in this field the term frequency matrix. One column corresponds to the frequencies of one term in each document. The selected data must be numerical. If the variable header has been selected, check that the "Column labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet in the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selection contains a label.

Document labels: Activate this option if the document labels are available. Then select the corresponding data. If the "Column labels" option is activated, you need to include a header in the selection. If this option is not activated, the document labels are automatically generated by XLSTAT (Doc1, Doc2, etc.).

Options tab:

Alpha: α corresponds to the elastic-net mixing parameter which is between 0 and 1. When $\alpha = 1$ it is the LASSO penalty that is applied, and when $\alpha = 0$ it is the Ridge penalty.

Lambda: Choose how to select the λ values to test during the cross-validation.

- **Automatic**: Select this option to generate automatic λ values.
  - **Number of lambda values**: Enter the number of λ values to generate. Default value: 100.
- **Custom lambda values**: Select this option to enter manually the λ values by selecting a unique column with as many rows as λ values.

Iterations: Enter the maximum number of iterations. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 10000.

Number of folds: Enter the number of folds to be constituted for the cross-validation. Default value: 10.

Maximum variables: Enter the maximum number of variables to be used in the model.

Prediction tab:

Term frequencies (Prediction): Activate this option if you want to select data to use in prediction mode. If you activate this option, you need to make sure that the prediction dataset is structured exactly like the estimation dataset: the same variables to be selected in the same order. If the variable header has been selected, check that the "Column labels" option has been activated.

Document labels (Prediction): Activate this option if document labels are available. Then select the corresponding data. If this option is not activated, the document labels are automatically generated by XLSTAT (PredDoc1, PredDoc2 …).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Outputs tab:

Select coefficients according to the: Select the coefficients according to the optimal λ of your choice. * **Lambda minimum**: Select this option to choose the coefficients according to the λ that gives minimum mean cross-validated error. * **Lambda 1se**: Select this option to choose the coefficients according to the λ that gives the most regularized model such that the cross-validated error is within one standard error of the minimum.

Optimal lambda: Activate this option to display a table with the values and the degrees of freedom according to the λ.

Coefficients: Activate this option to display the sorted coefficients of each term.

Odds ratio: Activate this option to display the odds ratio of each term in the same table as the coefficients.

Term frequencies: Activate this option to display the total term frequency of each term in the same table as the coefficients.

Display non-zero coefficients only: Activate this option to display only terms with an influence according to the model. Terms with zero coefficients, their odds ratio, and their frequency are then removed from the "Results by term" table.

Results by document: Activate this option to display the response variable and the prediction for each document and the probabilities for the classification.

Confusion matrix: Activate this option, only in the case of classification, to display the confusion matrix for the classification of the training dataset. The confusion matrix contains information about the observed and predicted classifications by the model. Performances can be evaluated using the confusion matrix. The diagonal contains correct predictions. The greater the sum of elements of the diagonal, the better the classifier.

Goodness of fit statistics: The statistics related to the fitting of the regression model are shown in this table.

Charts tab:

Coefficients: Activate this option to display a bar chart showing the term coefficients.

Odds ratio: Activate this option to display a bar chart showing the term odd ratio.

Evolution of the deviance: Activate this option to display a chart showing the cross-validation curve with its upper and lower standard deviation curves, as a function of the λ values automatically generated or entered (see Options tab). The λ minimum is plotted in red while the λ 1se is plotted in blue. If the two λ are equal, only the λ minimum is plotted.

Results

Results regarding the terms: This table gives a view of the influence of each term. The coefficient and the odd ratio allow you to know if a term is important or not. The coefficient gives the intensity and direction of its influence whereas the odd ratio gives the probabilities to predict the target class vs another target. For instance, if the target class is "Positive" and the other one is "Negative" and the odd ratio for the term "good" is three, it means that the document with "good" includes in it, presents the chance to be predict as "Positive" three times greater than a document which does not have this term. The frequency column helps to know if the coefficient is influenced by a high frequency. If no term has a non-zero coefficient only the intercept is plotted on the coefficients and odds ratio charts. To get more terms with a non-zero coefficient, we suggest decreasing the value of the α.

Results regarding the confusion matrix: The confusion matrix is deduced from prior and posterior classifications together with the overall percentage of well-classified observations.

Results regarding the goodness of fit statistics: The statistics related to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations.
- **DF**: The number of degrees of freedom for the chosen model.

- **Deviance**: Corresponds to the loss, for the Gaussian model it is the squared error, for the Poisson model it is the deviance and for binomial or multinomial classification it is the misclassification error.
- **AIC**: Akaike's Information Criterion.
- **AICc**: Corrected Akaike's Information Criterion.
- **SBC**: Schwarz's Bayesian Criterion.

Results regarding the documents: This table gives a view of the document prediction. For the classification case, the probabilities for the target class are displayed for binomial classification and the probabilities for each class are displayed for multinomial classification. Note: The target class is the last in alphabetical order.

Example

A tutorial on the terms selection feature is available at the XLSTAT Help Center:

https://www.xlstat.com/demo/trs_en

References

Hastie, T., Qian, J., & Tay, K. (2021). An Introduction to glmnet. CRAN R Repository.

Decision aid

Multicriteria decision aid: ELECTRE methods

In a decision-making process ELECTRE methods enable to identify a set of solutions to a problem, to compare solutions or to classify them from the best to the worst. These methods have the advantage of taking into account several criteria that may be of different nature (qualitative or quantitative) and of chosen importance order.

In this section:

Description

The ELECTRE methods, whose acronym stands for ELimination and Choice Expressing REality, bring together a family of decision aid methods whose particularity is the partial aggregation based on the construction of relations of comparisons of the performances of each pair of solutions. Unlike classical optimization methods, which consist in formulating the problem in the form of a cost function and in searching its optimum, here we compare solutions 2 by 2, criterion by criterion putting forward a preference / indifference of a response to another and resulting in an over ranking matrix. These methods have the advantage of accepting situations of incomparability with qualitative and immeasurable criteria.

The different solutions of a decision-making problem are called potential actions, or alternatives. These actions are listed exhaustively or not and must be formulated by the user. The consequences of each of them are evaluated using criteria. A criterion can be qualitative or quantitative and must be defined by the user. When it is qualitative, the evaluation of the actions on this criterion must be reduced to a numerical scale defined by the user. For example, let's consider the criterion "type of diploma" for the selection of a candidate in a recruitment process. This criterion will be reduced to an arbitrary numerical scale which can be: 0 for high-school diploma, 1 for associate degree, 2 for bachelor degree, and so on... To allow a different contribution of these criteria in the decisional problem, the user can provide a weight to each one increasing with its importance. In the example of the recruitment process let "age of the candidate" be a new criterion ranging from 25 to 50 years. Assuming that this criterion is less important than the criterion "type of diploma" then the weight value of the criterion "Age of the candidate" is set to 1 and the weight value of the criterion "type of diploma" is set to 2. In the

end, in order to use an ELECTRE method the user must provide at least: the list of actions, the list of criteria, the evaluation of each action by criterion and the weight of each criterion.

Let A be a finite set of p potential actions, A = {a_1, a_2, ..., a_p}. Let F be a coherent family of n criteria, F = {g_1, g_2, ..., g_n} where each function g, representing the evaluation of actions by criterion, is defined on A and takes its values in a totally ordered set. Let K be the set of weight values associated with each criterion, K = {k_1, k_2, ..., k_p}. The table named “performance” is composed of n lines and p columns.

A true-criterion is dissociated of a pseudo-criterion according to the accuracy of the evaluations. If they are easy to establish being free of error margins, we speak of true-criteria. Conversely, if the evaluations are vague and imprecise, we speak of a pseudo-criterion. In this second case, the computation method will be supplemented with discrimination thresholds in order to increase the realism of the preference modeling.

XLSTAT proposes two methods ELECTRE, 1 and 3. Above is given a description of them.

Electre 1

This method is used to identify a set of solutions to a decision-making problem. The criteria are true-criteria and in this case the thresholds p and q are set to 0. Let a and b be two potential actions, Electre 1 gives an over ranking matrix that numerically translates the assertions "a over ranks b", noted aSb, meaning that the action a is privileged over the action b and the opposite assertion. To do this, we need to compute 2 matrices, one called concordance matrix and the second one called discordance matrix.

The indexes of the concordance matrix for two actions a and b are denoted by C (a, b), ranging from 1 to 0, and measure the relevance of the assertion "a over ranks b" as follows: $\forall a, b \in A$,

$$C(a,b) = \frac{1}{\sum_{j=1}^{n} k_j} \sum_{j=1}^{n} k_{j/g_j(a) \geq g_j(b)} \qquad (1)$$

The indices of the discordance matrix are denoted by D (a, b), ranging from 1 to 0, and measure the relevance of an argument against the assertion "a over ranks b" as follows: $\forall a, b \in A$,

$$D(a,b) = \frac{1}{\delta} \max_{j=1 \text{à} n} (g_j(b) - g_j(a)), \qquad (2)$$

with

$$\delta = \max_{j=1 \text{à} n} (\max_{i=1 \text{à} p} (g_{j,i}(a) - g_{j,i}(b))).$$

The over-ranking matrix is constructed from all of these 2 indices via the following over-ranking relationship: $\forall a, b \in A$,

$$aSb \Leftrightarrow \begin{cases} C(a,b) \geq \hat{c} \\ D(a,b) \leq \hat{d} \end{cases} \qquad (3)$$

where $\hat{c}$ is the concordance threshold and $\hat{d}$ the discordance one. These thresholds must be taken in the interval [0,1]. When both inequalities of (1) are true then the index of over ranking is 1, otherwise it is equal to 0. We can thus deduce, for each action, the number of times it outperforms and the number of times it is outperformed. This result is summarized in a table classifying the actions according to the number of over rankings. Actions with the same number are ranked in the same rank. By default, the thresholds are set to 1 and 0 respectively but for more flexibility in the method and to allow to weaken the assertion aSb they can be varied by the user.

The method is supplemented by a sensitivity analysis on the thresholds. It enables to identify the minimum and maximum values for which the final result of the over ranking remains unchanged. Electre 1 is then runned 4 times: 2 times by modifying the concordance threshold and fixing the discordance threshold at the value provided by the user (or at the default value 0) and 2 times by fixing the concordance threshold at the value given by the user (or the default value of 1) and modifying the discordance threshold. The modified values are increased and decreased by 10% of the user value.

Electre 3

This method is used to classify a set of solutions from the best to the worst. Criteria are pseudo-criteria and in this case thresholds are required to do the analysis. Compared to Electre 1, Electre 3 executes more computations in order to obtain the desired results. In a first step the method computes matrix coefficients which summarize the information of concordance and discordance between actions of the problem. In a second step, the coefficients are used to build two pre-rankings, a first one which classifies solutions from the best to the worst and a second one which classifies from the worst to the best. The outranking matrix and the table rank are then deduced by crossing the two pre-rankings results.

Let q_j be the indifference threshold, p_j be the preferred threshold and v_j be the veto threshold such that $q_j < p_j < v_j$ over the criterion j. Let u_j = g_j(a) – g_j(b) be the difference between the performances of actions a and b over the criterion j. We compare u_j to the thresholds and define the outranking relations as follow:

- a and b are indifferent (a I b) $\Leftrightarrow u_j \leq q_j(g_j(a))$,
- a is less preferred to b (a R b) $\Leftrightarrow q_j(g_j(a)) \leq u_j \leq p_j(g_j(a))$,
- a is preferred to b (a P b) $\Leftrightarrow p_j(g_j(a)) \leq u_j$,
- a is less better than b (a NP b) $\Leftrightarrow u_j \geq v_j(g_j(a))$.

The aim of Electre 3 is to get an over ranking matrix that translates the assertions aSb with S = I, R, P and NP, for all couple of actions a and b in A. To do this, we need to compute the matrix coefficients according to the following equation:

$$d(a,b) = \begin{cases} C(a,b) & \text{if } \forall j \ D_j(a,b) > C(a,b), \\ C(a,b) \prod_{j=1}^{n} \frac{1-D_j(a,b)}{1-C(a,b)} \end{cases} \qquad (4)$$

where C(a, b) represents the index of the global concordance matrix and D_j(a, b) the index of the j-partial discordance matrix. These two matrices are respectively computed as follow. The global concordance between two actions a and b is a linear combination of the partial

concordances c_j(a,b), for j=1 to n, normalised by the weights of criteria as formulated in equation (5).

$$C(a,b) = \frac{\sum_{j=1}^{n} k_j \times c_j(a,b)}{\sum_{j=1}^{n} k_j} \qquad (5)$$

with

$$c_j(a,b) = \frac{p_j(g_j(a)) - min\left(g_j(b) - g_j(a), p_j(g_j(a))\right)}{p_j(g_j(a)) - min\left(g_j(b) - g_j(a), q_j(g_j(a))\right)}$$

This calculation involves the distance u_j, the preference and indifference thresholds. The computation of the partial discordance matrices (one matrix per criterion) also involves the distance u_j and the preference threshold but also the veto threshold with the following equation.

$$D_j(a,b) = Min\left(1; Max\left(0; \frac{g_j(b) - g_j(a) - p_j(g_j(a))}{v_j(g_j(a)) - p_j(g_j(a))}\right)\right).$$

The credibility matrix is then used in an iterative algorithm which is executed two times. A first time to find the best solutions to the worst ones, this step is called descending distillation. And a second time to get the inverse ranking, that is to say from the worst solutions to the best, this step is called ascending distillation. The algorithm aims at the iteration i to extract a subset Ai of A composed of the best or the worst solutions depending on the distillation step followed. The inclusion rule relies on the comparison with the credibility matrix coefficients to the discrimination threshold that is computed with the following equation:

$$s_i(\lambda) = \alpha + \lambda_i \times \beta,$$

where α = 0.30, β = – 0.15 and λ_i represents the biggest value of all degrees of credibility.

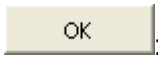
The expected outranking matrix is then deduced by crossing the two distillation steps. The final ranking is deduced from the number of occurrences I, R, P and NP, whose the meaning is given above, of the outranking matrix.

To allow method flexibility, several options are proposed to the user. The first one enables to choose the preferred direction of performances for each criterion. An increasing direction means that the performances increase and the preference is given to the high value (maximisation of the criterion). A decreasing direction means that the performances decrease and the preference is given to the low value (minimization of the criterion). This option is traduced in XLSTAT by a value of 1 for the maximization and -1 for the minimization. A default value is prescribed that is 1. The next two options are linked with the format and the direction of the thresholds. Those can be defined as a constant or as a linear function of the performance g_j(a). The linear function is chosen when the difference uj is large. In this case the thresholds are defined either in direct mode, that is to say the performance used in the threshold computations is the worst one, or in inverse mode that is the performance used is the better one. This option is traduced in XLSTAT by a value 1 for the direct mode and -1 for the inverse mode. A default value is prescribed that is 1.

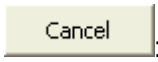
To use Electre 3 method, the user must provide in addition to the elements common to both methods mentioned above, the indifference, preference and veto thresholds, and to choose the threshold format.

Dialog box

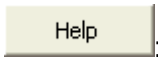
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Criteria/actions table: You can select a table that contains the performances of the actions. The table size must be of N lines and P columns and data must be of numerical type. If column headers have been selected, check that the "Variable labels" option has been activated.

Criteria weights: You can select a table that contains the weight of each criteria. This table must have the same number of lines than the criteria/actions table and data must be of numerical type. If a column header has been selected, check that the "Variable labels" option is activated. Missing data are not accepted.

Type of the method: You can choose between Electre 1 and 3 methods for your computations (see the description section for more details).

Concordance threshold: Threshold used in Electre 1 method. You can choose the value of the concordance threshold, between 0 and 1. The default value is 1. This threshold must be greater than the discordance one.

Discordance threshold: Threshold used in Electre 1 method. You can choose the value of the discordance threshold, between 0 and 1. The default value is 0. This threshold must be smaller than the concordance one.

Indifference threshold: Threshold used in Electre 3 method. You can select a table that contains the indifference threshold of each criteria. This table must have the same number of

lines than the criteria/actions table and 1 or 2 columns depending on the format chosen (see the selection «Threshold format » below). Data must be of numerical type. If a column header has been selected, check that the "Variable labels" option is activated. Missing data are not accepted.

Preference threshold: Threshold used in Electre 3 method. You can select a table that contains the preference threshold of each criteria. This table must have the same number of lines than the criteria/actions table and 1 or 2 columns depending on the format chosen (see the selection «Threshold format » below). Data must be of numerical type. If a column header has been selected, check that the "Variable labels" option is activated. Missing data are not accepted.

Veto threshold: Threshold used in Electre 3 method. You can select a table that contains the veto threshold of each criteria. This table must have the same number of lines than the criteria/actions table and 1 or 2 columns depending on the format chosen (see the selection «Threshold format » below). Data must be of numerical type. If a column header has been selected, check that the "Variable labels" option is activated. Missing data are not accepted.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (row and column variables, weights) includes a header.

Criteria labels: You can select a table that contains the criteria labels. This table must have the same number of lines than the criteria/actions table and data must be of character type. If you do not select this table, the default labels are "Crit.1, Crit.2, …". If column headers have been selected, check that the "Variable labels" option has been activated.

Threshold format: Option used in Electre 3 method. You can choose the format of thresholds between constant and linear. In the constant case a table of the same number of lines than the criteria/actions table and 1 columns is expected. In the linear case two columns are expected, the first one corresponding to the slope and the second one to the intercept.

Options tab:

Criteria evaluation direction: You can activate this option to add a condition on the direction of evaluation of each criteria. You select a table that contains the same number of lines than the criteria/actions table and 1 column. Data must be of numerical type set to 1 or -1. By default, data are fixed to 1. If a column header has been selected, check that the "Variable labels" option is activated. Missing data are not accepted.

Threshold direction: You can activate this option to add a condition on the direction of each tresholds. You select a table that contains the same number of lines than the criteria/actions table and 1 column. Data must be of numerical type set to 1 or -1. By default, data are fixed to 1. If a column header has been selected, check that the "Variable labels" option is activated. Missing data are not accepted.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove criteria: Activate this option to remove criteria with missing data.

Remove actions: Activate this option to remove actions with missing data.

Estimate missing data: Activate this option to estimate the missing data before the calculation starts.

- **Mean**: Activate this option to estimate the missing data by using the mean (quantitative variables) of actions.
- **Nearest neighbor**: Activate this option to estimate the missing data for an criterion by searching for the nearest neighbor of it.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the actions selected. For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed.

Concordance matrix: Activate this option to display the concordance matrix. For Electre 3 method the global matrix is displayed.

Discordance matrix: Activate this option to display the discordance matrix. Option only available for Electre 1 method.

Credibility matrix: Activate this option to display the credibility matrix. Option only available for Electre 3 method.

Over ranking matrix: Activate this option to display the over ranking matrix.

Ranking table: Activate this option to display the table of ranks.

Sensitivity analysis: This option is allowed if « Ranking table » is checked in Electre 1 method. It displays the ranking tables obtained with a sensitivity analysis of 10% on the thresholds.

Results

XLSTAT displays a large number tables and charts to help in analyzing and interpreting the results.

Summary descriptive statistics: the table of descriptive statistics displays the number of observations, the number of missing observations, the minimum, the maximum, the average, and the unbiased standard deviation of each action.

Concordance matrix: This result displays the indexes of the concordance matrix computed with the equation (1) for Electre 1 method, and equation (5) for the Electre 3 method (see the

description section).

Discordance matrix: This result displays the indexes of the discordance matrix computed with the equation (2) given in the description section of the method Electre 1.

Credibility matrix: This result displays the indexes of the credibility matrix computed with the equation (4) given in the description section of the method Electre 3.

Over ranking matrix: This result displays the matrix of 0 and 1 got with the over ranking relations (3) given in the description section of the method Electre 1. For Electre 3 method characters I, R, P and NP are displayed.

Ranking table: This result displays a table with the final rank of actions.

Ranking tables of the concordance threshold sensitivity analysis: This result displays 2 tables with the final rank of actions obtained with Electre 1 using a modified concordance threshold and a discordance threshold fixed to the user value (or set to the default value 0). The left table is the result with a 10% increase of the user value and the right table is the result with a 10% decrease.

Ranking tables of the discordance threshold sensitivity analysis: This result displays 2 tables with the final rank of actions obtained with Electre 1 using a modified concordance threshold and a discordance threshold fixed by the user (or set to the default value 0). The left table is the result with a 10% increase of the user value and the right table is the result with a 10% decrease.

Example

A tutorial on Electre 1 method is available on the XLSTAT Help Center:

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

A tutorial on Electre 3 method is available on the XLSTAT Help Center:

http://www.xlstat.com/demo-elc3.htm

References

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.

Design of experiments for the analytic hierarchy process

Use this tool to generate experimental designs needed to run analytic hierarchy process (AHP) analysis.

In this section:

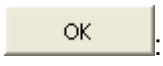
Description

The principle of AHP method is to compare the solutions of a decision problem using on a set of criteria in order to deduce the best solution. The application of such method requires defining several numerical tables whose number can quickly increase according to the characteristics of the decision problem (number of criteria, of sub criteria, of alternatives and of evaluators). In a simple case of 4 alternatives, 4 criteria and 2 evaluators, 9 tables of size 4x4 must be user defined and provided to allow the computations of the AHP method. Each table is a square matrix with 1 on the diagonal and Saaty's table values in the other elements of the table (see the help document on the AHP method).

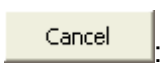
To limit typing errors XLSTAT offers the DHP tool to automatically generate tables useful in the AHP analysis. Its use is simply by specifying the characteristics of the decision problem, that is the list of alternatives, the list of criteria and sub-criteria if they exist, and the list of evaluators if necessary (see description of the AHP method for the definition of terms).

Dialog box

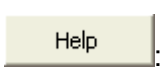
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: click this button to start the calculations.



: click this button to close the dialog box without doing any calculations.



: click this button to display the help.


: click this button to reload the default options.


: click this button to delete the data selections.

 : click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Alternatives: select a table that contains the list of alternatives labels. The data must be of type character. If a column header has been selected, check that the "Variable labels" option is activated. Missing data are not accepted.

Criteria: select a table that contains the list of criteria labels. The data must be of type character. If a column header has been selected, check that the "Variable labels" option is activated. Missing data are not accepted.

Subcriteria: activate this option if you want to select a table that contains the list of sub criteria labels. The data must be of type character. This table must have the same number of columns than the number of criteria. If a column header has been selected, check that the "Variable labels" option is activated. Missing data are not accepted.

Evaluator number: enter the total number of evaluators.

Evaluator Labels: activate this option if you want to select a table that contains the list of evaluator labels. The data must be of the type string. This table must have the same number of columns than the number of evaluators. If this option is not checked the evaluator label is numerical. If a column header has been selected, check that the "Variable labels" option is activated. Missing data are not accepted.

Range: activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: activate this option to display the results in a new worksheet of the active workbook.

Workbook: activate this option to display the results in a new workbook.

Variable labels: activate this option if the first row of the data selections (alternatives, criteria, subcriteria, evaluator) includes a header.

Results

Summary table: this table summarizes all the selected data. This table displays in this order the list of criteria, of subcriteria if the option is checked, of alternatives and of evaluators if the option is checked.

Saaty Table: this table displays Saaty's values, a definition of each value, and comments. Below the table some sentences explain how to use Saaty's values.

Comparison matrix of the evaluator X: this output displays on 2 or 3 lines of tables the comparison matrices that must be defined by the evaluator x. The criteria comparison matrix is displayed on the first table row. Just below, the comparison matrix of the sub-criteria is

displayed if the "subcriteria" option is selected. Finally, the alternative comparison matrix is displayed if the "sub-criteria" option is checked, otherwise they are displayed on the second table row.

Criteria comparison matrix: this output, in table form, displays the criteria comparison matrix. The cells on the diagonal are set to 1 and the others are empty. The cells below the diagonal are blocked. Only cells above the diagonal can be entered with values from Saaty table (see description section).

Subcriteria comparison matrix: this output displays, in the form of one or more tables on the same line, the subcriteria comparison matrix if the "sub-criteria" option has been checked. The cells on the diagonal are set to 1 and the others are empty. The cells below the diagonal are blocked. Only cells above the diagonal can be entered with values from Saaty table (see description section).

Alternative comparison matrix: this output displays, in the form of several successive tables on the same line, the comparison matrix of alternatives according to the criteria and the selected sub-criteria The cells on the diagonal are set to 1 and the others are empty. The cells below the diagonal are blocked. Only cells above the diagonal can be entered with values from Saaty table (see description section).

Example

A tutorial on DHP tool is available on the XLSTAT Help Center:

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

Multicriteria decision aid: AHP method

Use this application to solve your decision problem based on Analytic Hierarchy Process (AHP). In order to use this method, you must generate a design of experiments with the XLSTAT DHP tool also available under the XLSTAT menu Decision Aid.

The AHP method is a decision aid method based on a criteria hierarchisation. It is better adapted when the criteria number remains reasonable, and when the user is able to evaluate 2 by 2 the elements of his problem. The AHP feature proposed in XLSTAT has the advantage of not having any limitations on the number of criteria, of subcriteria and of alternatives and allows the participation of a large number of evaluators.

In this section:

Description

Analytic Hierarchy Process is a method adapted to multi-criteria decision problems that have several solutions satisfying a set of criteria. The approach of the method is to simplify the problem by breaking it down into a hierarchical system. Thomas Saaty is at the origin of this method and created it in the 1970s.

We denote by *alternative* the solutions of the decision problem, *criteria* the parameters on which the alternatives are evaluated, *subcriteria* the parameters belonging to a criterion and on which alternatives are evaluated and *evaluator* the person who makes the evaluations. We talk about a 2-level problem when it admits subcriteria, otherwise, it is a problem of level 1.

The principle of the method is to evaluate 2 by 2 the elements of the problem in comparison tables. There are tables to define at each level of the hierarchy. At level 0 the criteria comparison table must be defined, at level 1 those of subcriteria if it is a level 2 problem, otherwise the comparison tables of alternatives over criteria. At level 2 the comparison tables of alternatives over criteria and / or subcriteria must be defined. The set of these tables is called the experimental design for an AHP analysis.

XLSTAT proposes the DHP tool to create your own design of experiments. It is available in the Decision Aid menu under the Advanced features ribbon. A description of the DHP is available here.

The comparison tables must be completed by the user according to the values chosen in the table of Saaty reported below. Saaty has defined an evaluation scale that measures the importance or the difference of one element over another one.

Definition	Value
Equal importance of an element over another one	1
Moderate importance of an element over another one	3
Strong importance of an element over another one	5
Very strong importance of an element over another one	7
Extreme importance of an element over another one	9
can be used to express intermediate values	2, 4, 6, 8
Reciprocity	1/above value

The first computation of the AHP method is the calculation of the criteria priority vector from the values of the criteria comparison table, i.e. the computation of the weight of each criterion. The equation is given by:

the weight of each criterion = sum of normalized lines/criteria number.

With the same equation the priority vector of subcriteria is calculated for each criterion. This vector is then weighted by the weight of the criterion linked. We thus obtain the weights of each criterion and subcriterion which are finally used as the weighting in the calculation of alternatives priority vectors with the same mathematical equation given above.

An option in the output results is proposed to assess the data consistency. This test controls the entered values in the comparison tables. Indeed, if the alternative A1 is evaluated slightly more important (code 2) than the alternative A2, and A2 is judged moderate important (code 3) than the alternative A3 and A3 strong important (code 5) than A1, then the test says that there is a data inconsistency. It is measured with 2 parameters: the coherence index (CI) and the coherence ratio (CR). The equation for the coherence index is as follows.

CI = mean coherence – element number / (element number -1)

The number of items is the number of columns or rows in the comparison table. To obtain the average consistency, we first multiply the comparison matrix with its priority vector, which gives us a new vector. Then, we divide the latter by the weight of the priority vector of the element of the same line. The average of this normed vector gives the average coherence. The formula for calculating the consistency ratio is given by:

CR = coherence index / random coherence

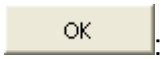
where the random coherence is given by the following table:

Criteria number	2	3	4	5	6	7	8	9
Random coherence	0.00	0.58	0.90	1.12	1.24	1.32	1.41	1.45

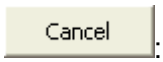
If the coherence ratio is less than or equal to 10% then the assessment is considered consistent. On the other hand, if it is larger than 10%, it is recommended to review the evaluation of the concerned comparison table.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: click this button to start the calculations.



: click this button to close the dialog box without doing any calculations.

: click this button to display the help.


: click this button to reload the default options.


: click this button to delete the data selections.


: click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Experiment plan sheet: select a sheet containing the XLSTAT generated DHP design (see help of the DHP method) to run an AHP analysis.

Range: activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: activate this option to display the results in a new worksheet of the active workbook.

Workbook: activate this option to display the results in a new workbook.

Variable labels: activate this option if the first row of the data selections (alternatives, criteria, subcriteria, evaluator) includes a header.

Missing data tab:

Do not accept missing data: activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Replace with the equal importance value 1: activate this option to replace the missing data in a comparison table with the importance value 1 of the Saaty table (see description section above). This gives equal importance between the two elements of comparison.

Outputs tab:

Mean table of results: activate this option to display mean table or tables of results got by evaluator. This option is available if there are at least 2 evaluators.

Mean table of criteria: activate this option to display the mean table of criteria got by evaluator. This option is available if there are at least 2 evaluators.

Mean table of subcriteria: activate this option to display the mean table of subcriteria got by evaluator. This option is available if there is at least 2 evaluators and selected subcriteria.

Mean table of alternatives: activate this option to display the mean table of alternatives got by evaluator. This option is available if there are at least 2 evaluators.

Results per evaluator: activate this option to display table or tables of results got for each evaluator.

Criteria: activate this option to display tables of results on criteria got for each evaluator.

Subcriteria: activate this option to display tables of results on subcriteria got for each evaluator.

Alternatives: activate this option to display tables of results on alternatives got for each evaluator.

Data consistency: activate this option to display the result of data consistency compute. This option is available if the option “results per evaluator” is activated.

Coherence index: activate this option to display the result of coherence index compute (see description section for more details). This option is available if the option “results per evaluator” is activated.

Coherence ratio: activate this option to display the result of coherence ratio compute (see description section for more details). This option is available if the option “results per evaluator” is activated.

Graphs tab:

Bar charts: activate this option to display table or tables of results got per evaluator as bar charts. This option is available if table results are requested.

Criteria: activate this option to display tables of results got on criteria per evaluator as bar charts. This option is available if the option “Bar charts” is activated and criteria table results are requested.

Subriteria: activate this option to display tables of results got on subcriteria per evaluator as bar charts. This option is available if the option “Bar charts” is activated and subcriteria table results are requested.

Alternatives: activate this option to display tables of results got on alternatives per evaluator as bar charts. This option is available if the option “Bar charts” is activated and alternatives table results are requested.

Results

Mean priorities by criterion: this result, given as a table, corresponds to mean relative percentages of weight vectors by criterion over all evaluators. If the options “Bar charts” and “criteria” are selected the result is displayed as charts under the table.

Mean priorities by subcriterion: this result, given as a table, corresponds to mean relative percentages of weight vectors by subcriterion over all evaluators. If the options “Bar charts” and “subcriteria” are selected the result is displayed as charts under the table.

Mean priorities by alternatives: this result, given as a table, corresponds to mean relative percentages of weight vectors by alternatives over all evaluators. If the options “Bar charts” and “alternatives” are selected the result is displayed as charts under the table.

Results obtained from the ratings of evaluator x: give the set of tables and bar charts obtained for evaluator x according to selected options:

- **Priorities by criterion**: are the relative percentages for each criterion.
- **Priorities by subcriterion of criterion XXXX**: are the relative percentages for each subcriterion of the criterion XXXX.
- **Priorities by alternatives**: are the relative percentages for each alternative.

If the option « data consistency » is activated the parameters CI and CR are computed and displayed under tables (see the section description for more details).

Example

A tutorial on AHP tool is available on the XLSTAT Help Center:

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

References

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

Decision trees

The Decision Tree feature in XLSTAT is a decision support tool that presents its final result as a set of choices in the graphic form of a tree. Different possible decisions are represented by nodes, or branches, and the “leaves,“ located at the ends of branches, represent the possible outcomes of the decisions made at each stage. Widely used in various fields, the decision tree’s main function is to determine an optimal path according to configurable criteria.

Dans cette section :

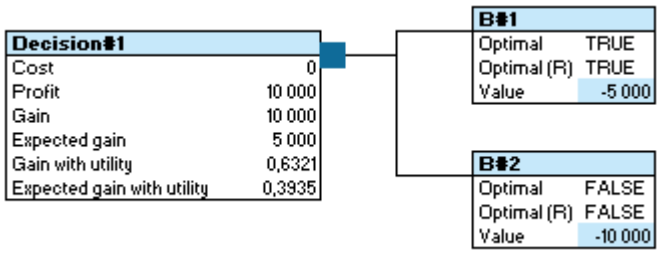
Description

A decision tree is a diagram represented by a set of nodes interconnected through branches. It allows its creator to evaluate different possible actions according to their cost, benefit and probability. It usually begins with a node from which several possible outcomes arise. Each of these results leads to other nodes, or children, from which other possibilities emanate. The resulting diagram recalls the shape of a tree.

Different types of nodes

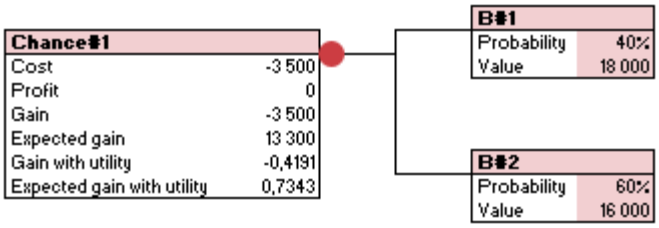
Here is a brief description for each type of node, with their graphical representation, in block form, in XLSTAT:

- **Decision node**: Represented by a blue square, it illustrates a decision to be made among several possible choices (branches). Here is an example of a decision node.

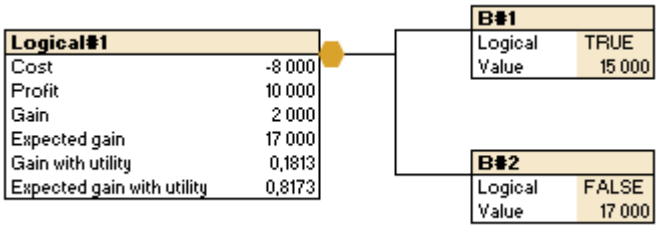


- **Chance node**: This is represented by a red circle. It offers different outcomes, each with a % of chance of occurrence. The different % of chance must sum to 100. Here is an

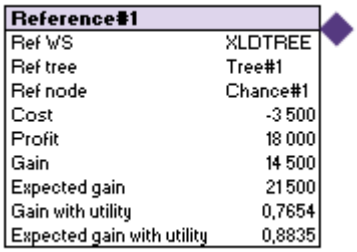
example of a chance node.



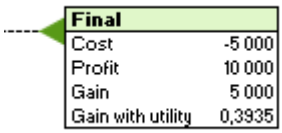
- **Logical node**: Represented by a yellow hexagon, this imposes a result or not according to the return of a logical formula. Here is an example of a logical node.



- **Reference node**: This is represented by a purple rhombus. It refers to a subtree, that is, to another node with all of its children. Here is an example of a reference node.



- **End node**: This is represented by a green triangle. It is the final result of a decision path. Here is an example of an end node.



Information available for each node

As you can see in the previous figures, each node includes different information that is available to you. You can choose to display all of this information or not. This setting will be discussed in the Building a tree section. In the meantime, here is a description of these different pieces of information. The concepts of calculation mode and utility function will be presented next.

- **Cost**: The cost of a node is the sum of all the costs of its parent nodes.

- **Profit**: The profit of a node is the sum of all the profits of its parent nodes.
- **Gain**: The gain of a node is the sum of its cost and its profit. It is therefore calculated according to the costs and profits of its parent nodes. The cost is a negative value and the profit a positive value.
- **Expected Gain**: The expected gain of a node depends on the gain of its child nodes. It helps to make a decision according to the chosen calculation method.
- **Gain with utility**: The ‘gain with utility’ of a node is calculated using the exponential utility function applied to the gain of this same node.
- **Expected Gain with utility**: The ‘expected gain with utility’ of a node depends on the ‘gain with utility’ of its child nodes. It helps to make a decision according to the chosen calculation method.
- **Referent leaf**: Only in the case of a reference node, it is the name of the leaf where the referent node is located. If this sheet is in a different workbook than the one where the current tree is located, then the name of this workbook is also filled in.
- **Referent tree**: Only in the case of a reference node, it is the name of the tree containing the referent node.
- **Referent node**: Only in the case of a reference node, it is the name of the referring node.

Computation modes

There are two calculation methods to help you in making a decision. You can choose to maximize your gain if you want to maximize your profit, or to minimize your gain if you want to optimize your cost. These two calculation methods are enhanced with the possibility of having a gain, the sum of costs and profits, but also a gain calculated from an exponential utility function.

In a risky universe, a perfectly rational individual makes investment decisions by maximizing the expectation (in the probabilistic sense of the term) of his utility function. This reflects the satisfaction generated by a given future wealth. In our case, we use the exponential utility function defined as follows:

$$U(x) = \frac{1 - \exp(-Rx)}{R}, R \neq 0$$

$$U(x) = x, R = 0$$

where x represents the gain and R the utility or degree of risk-aversion with $R > 0$ in the case of risk-aversion, $R = 0$ in the case of risk-neutrality and $R < 0$ if risk is sought. It is easy to see that a utility $R = 0$ amounts to having a gain with utility equivalent to a gain without utility.

Below we describe the calculations performed according to the type of node. We start with the final node because the gain, with utility or not, expected from a node depends on this same information for a child node.

- **End node**: In the case of the end node, the expected gain corresponds to the gain, and the expected gain with utility corresponds to the gain with utility.
- **Decision node**: the gain, with utility or not, depends on the method of calculation.
    - **Maximize gain without utility function**: The expected gain corresponds to the greatest expected gain among the expected gains of the direct child nodes.
    - **Maximize gain with utility function**: The expected gain with utility corresponds to the greatest gain with expected utility among the expected utility gains of direct child nodes.
    - **Minimize gain without utility function**: The expected gain corresponds to the smallest expected gain among the expected gains of the direct child nodes.
    - **Minimize gain with utility function**: The expected utility gain is the greatest expected utility gain among the expected utility gains of direct child nodes. In this case, the utility function becomes:

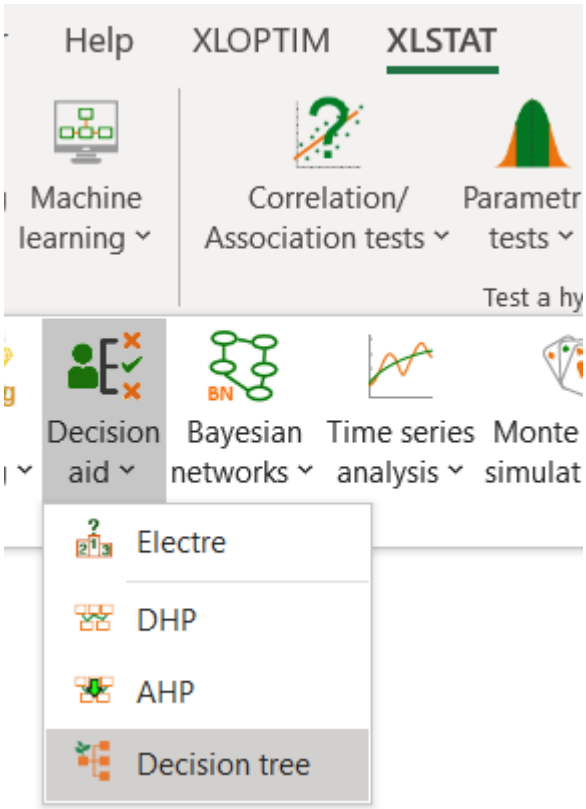
$$U(x) = \frac{1 - \exp(Rx)}{R}, R \neq 0$$

$$U(x) = x, R = 0$$

- **Chance node**: The expected gain, with utility or not, does not depend on the method of calculation. For each direct child node, it suffices to multiply its expected gain by its probability of realization. The expected gain of the parent node is thus the sum of these multiplications. The operation is the same in the case of the expected gain with utility.
- **Logical node**: The expected gain, with utility or not, does not depend on the method of calculation. It is that of the only child node whose logical formula is verified. If no child node has its logical formula verified then no expected gain, useful or not, can be calculated. In no case will this node be part of the possible decisions.
- **Reference node**: A reference node has a decision, chance or logic node as its referent. It is thus enough to refer to these types of nodes to know the calculation that is carried out.

Toolbar

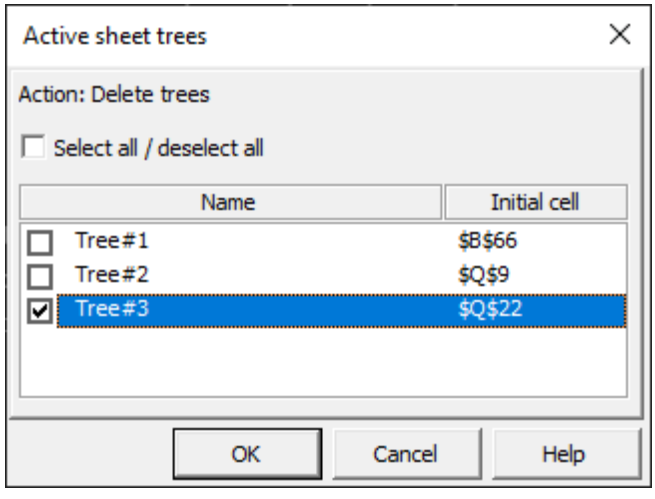
To build a new tree, you must first display the menu for decision trees. To do this, launch XLSTAT and find the Decision Tree tool in the ribbon as shown in the figure below:



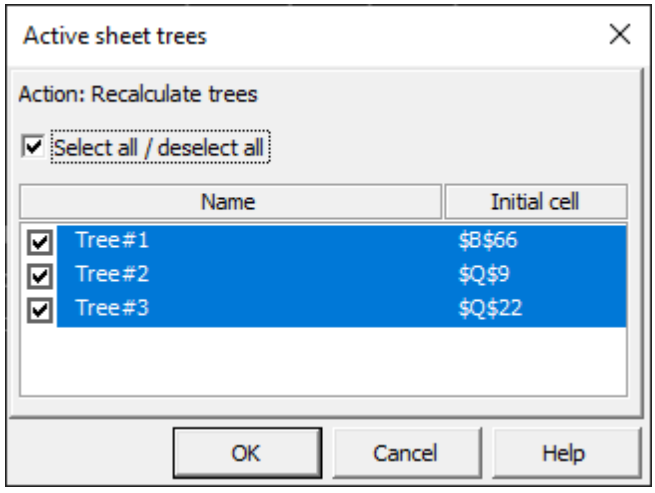
The menu shows as displayed below:

The different buttons that you can use are:

- **Create a new tree**: Click this button if you want to display the dialog box that will let you create a new tree.
- **Delete trees**: Click this button if you want to delete trees from the active sheet. A dialog box opens so you can select the trees to delete.



- **Recalculate trees** : Click this button if you want to recalculate trees in the active sheet. A dialog box opens so that you can select the trees to recalculate. This tool is useful when you have multiple trees in different tabs. It is possible that, when you modify the active tab, the result of the formulas of the trees present in this one are not up to date..



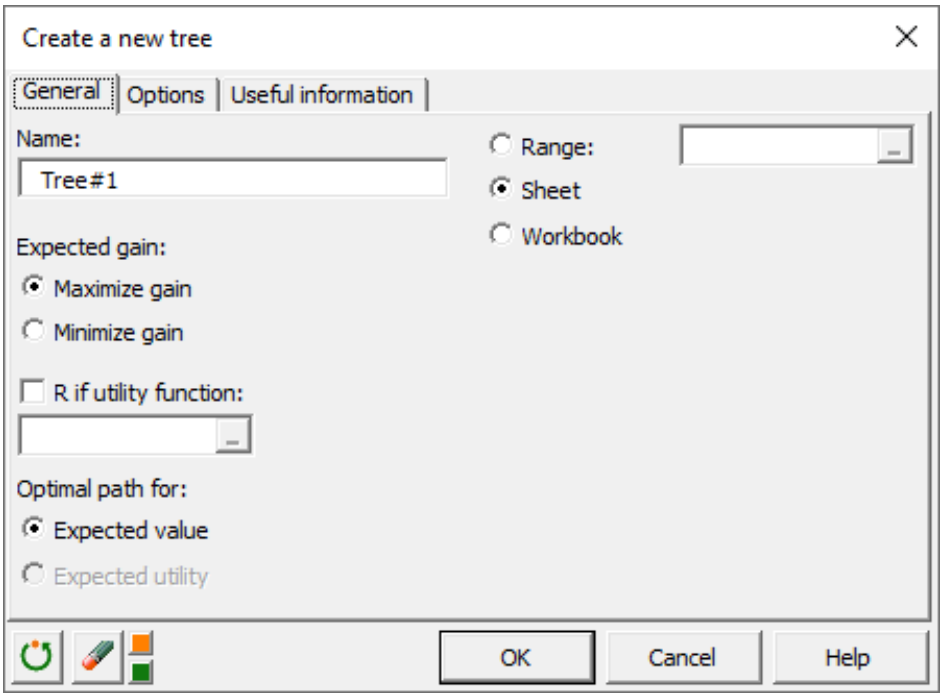
- **Hide gridlines**: Click this button if you want to hide the gridlines on the active sheet (only present if the grid is displayed).

- **Show gridlines** : Click this button if you want to show the gridlines on the active sheet (only present if the grid is not displayed).

Creating a tree

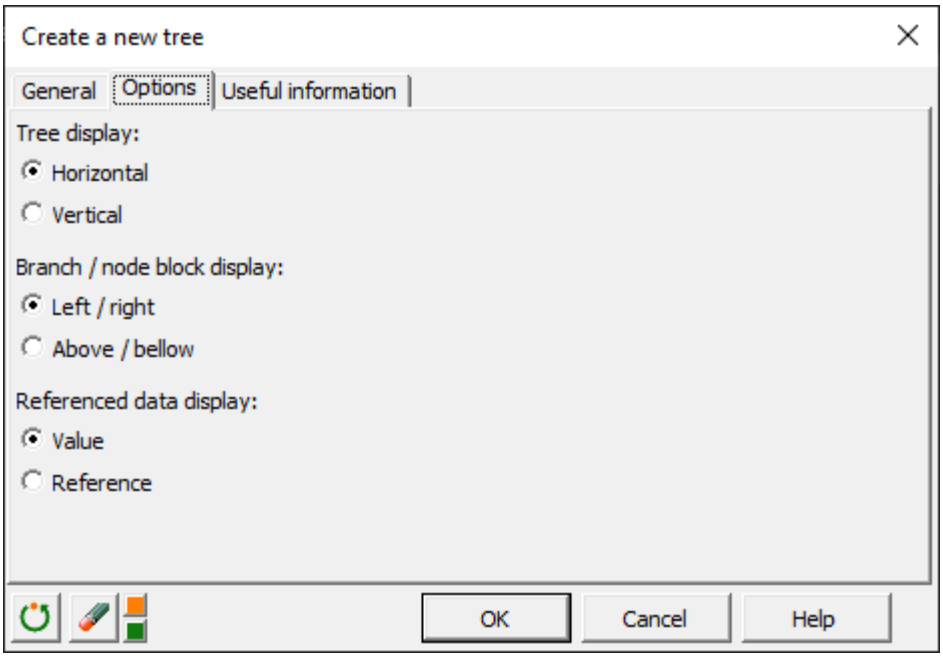
To create a new tree, all you have to do is click the corresponding button in the toolbar presented above. This displays a the dialog box where you can configure the tree. Below we will describe each of the available parameters. This dialog box can be opened at any time and you can modify the choices whenever needed. The tree updates automatically when you modify options.

General tab:



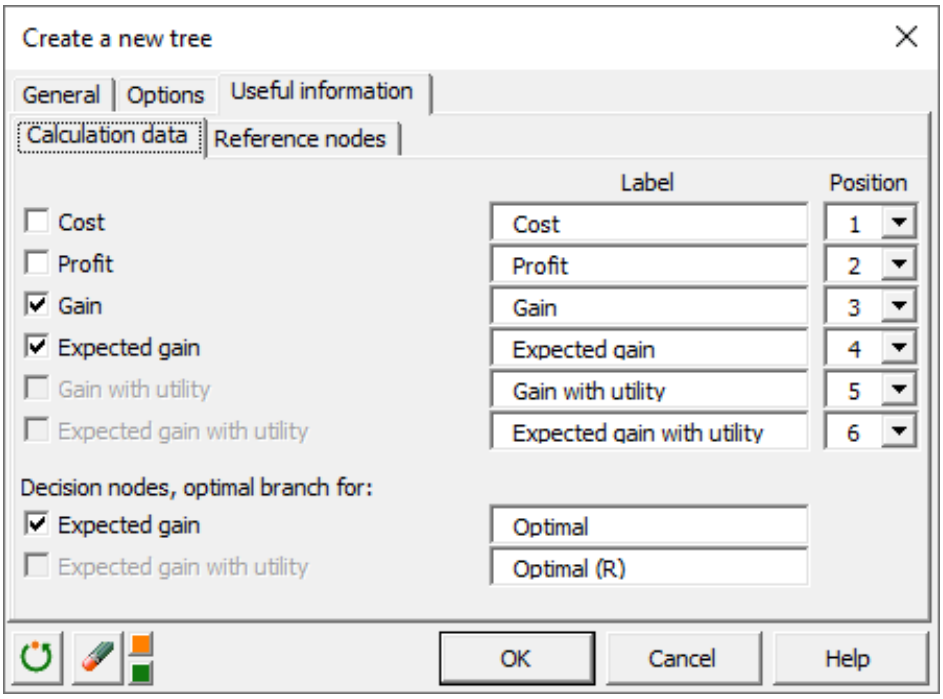
- **Name**: Enter the tree name.
- **Range**: Select this option if you want to select the starting cell of your tree yourself. As you build the tree, this cell corresponds to the cell at the top left of the smallest block containing your entire tree.
- **Sheet**: select this option if you want your tree to be created in a new sheet in the active workbook. The toolbar will be automatically added to this new sheet.
- **Workbook**: Select this option if you want your tree to be created in a sheet in a new workbook. The toolbar will be automatically added to this sheet.
- **Expected gain**: select your calculation method here. Possible options are **Maximize gain** or **Minimize gain**.
- **R if utility function**: Enter in this field your risk-aversion value (R value). You can also select a cell that contains the R value.
- **Optimal path for**: Select the optimal path type you want to display. If the option **Expected value** is selected then the optimal path will be calculated from the calculated expected gains. If the option **Expected utility** is selected, then the optimal path will be calculated from the calculated expected utility gains. This option can only be selected if **R if utility function** is also selected. *This setting is not available on Mac for the moment*.

Options tab:



- **Tree display**: Select the display mode for the tree. A **horizontal** tree expands from left to right. A **vertical** tree expands from top to bottom.
- **Branch / node block display**: Select the display mode for a branch block compared to a node block. The branch being the parent branch of the node block. A **Left / Right** display will display the node block to the right of the branch block. An **Above / Below** display will display the node block below the branch block.
- **Referenced data display**: Select the default display mode for the referenced data. Your choice is applied in the configuration interface of a node, when a data can refer to a cell in a sheet. Here you choose the default display mode, but it can still be modified in the configuration interface of a node.

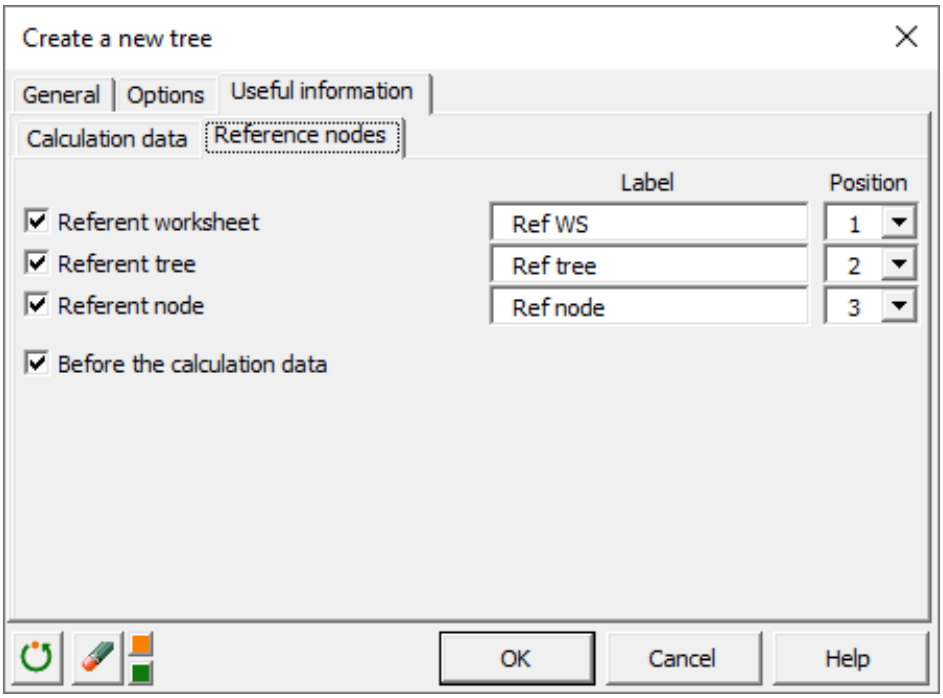
Useful information - Calculation data tab:



- Choose here the information to display for each node block. The definitions can be found in the Description section. For a given piece of information, it is possible to modify the associated label and its position in relation to the other information to be displayed. The information relating to calculations with utility can only be selected if the option **R if utility function** in the **General** tab is also selected.

- **Decision nodes, optimal branch for**: For decision nodes, at branch block level, it is possible to indicate whether or not the branch is optimal for the expected gain and / or for the gain with expected utility.

Useful information - Reference nodes :



- Choose the referent node information to display. The definitions can be found in the Description section. For a given piece of information, it is possible to modify the associated label and its position in relation to the other information to be displayed.

- **Before the calculation data**: This option allows you to choose, for each reference node, whether the information relating to the referring node is displayed before or after the calculation data.

When all options have been set:

OK: Click this button to validate your parameters and start building a new tree or updating the parameters of an existing tree.

Cancel: Click this button to close the dialog box without creating a new tree or modifying the parameters of an existing tree.

Help: Click on this button to display the XLSTAT decision trees help.

: Click this button to restore the default options.

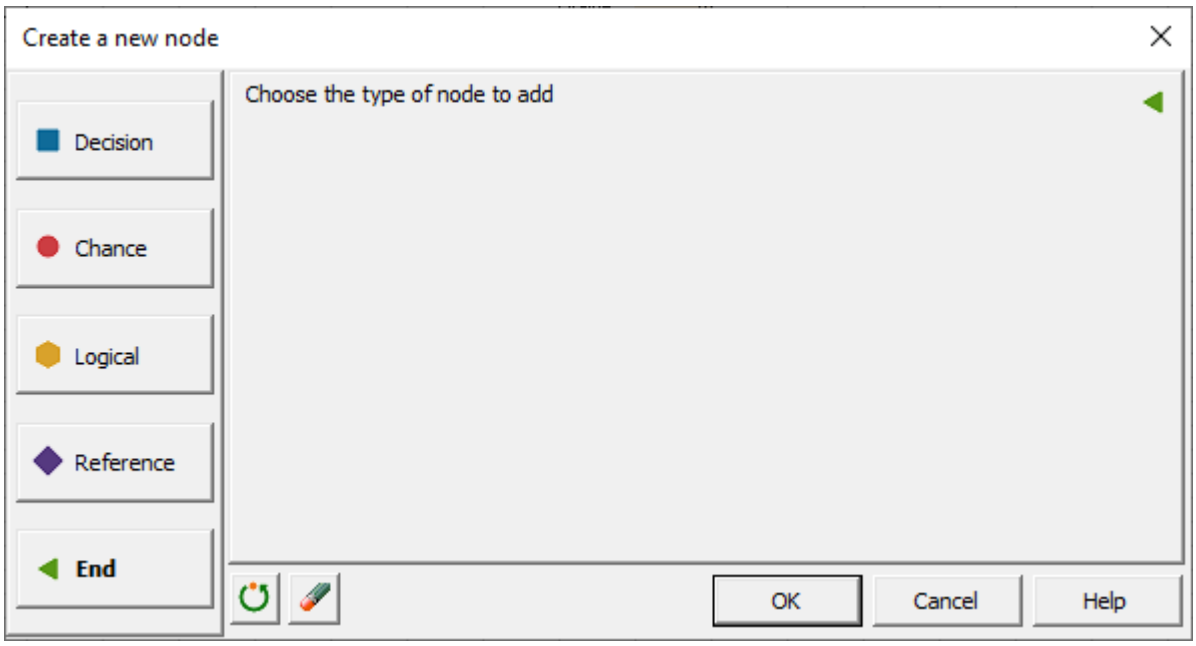
: Click this button to clear data selections.

: Click the orange button to save the dialog box settings to a file, or the green button to load the dialog box settings from a file.

Creating a node

To add a new node or modify an existing node, all you have to do is click on a node icon. This opens the dialog box for configuring a node. This dialog can be viewed and edited at any time. The target node will update automatically if you make any changes.

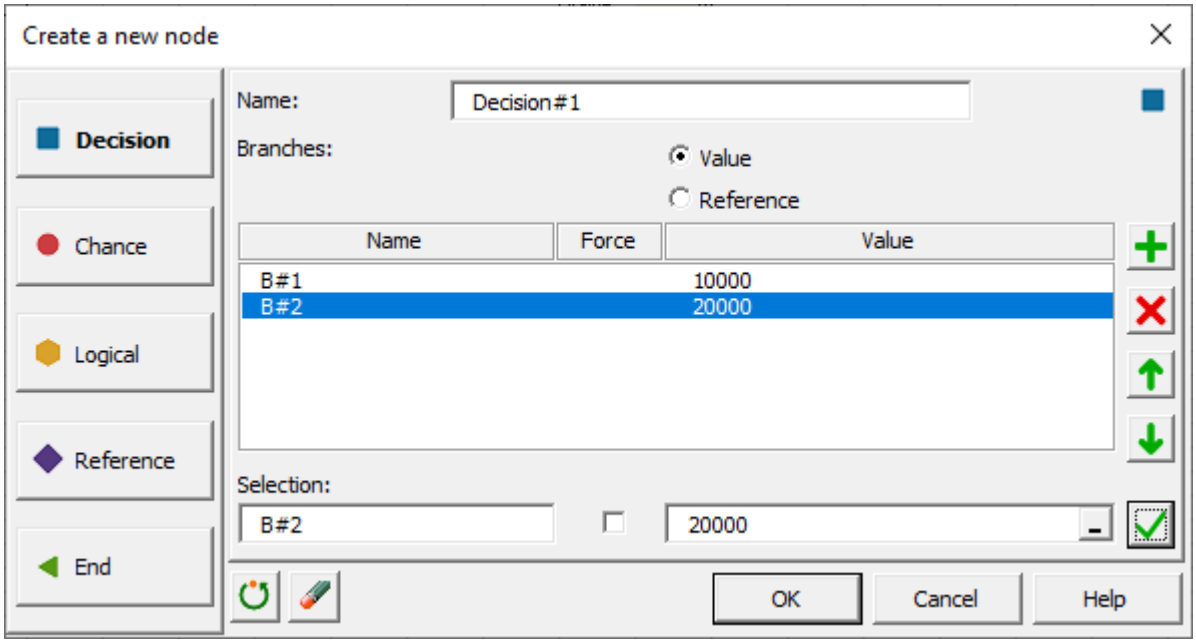
When you click the icon of an end node, a dialog box is displayed so that you can select the type of the new node.



Once the type is chosen, the dialog box expands and allows you to configure your node. This same extended dialog box appears when you click the icon of an existing node, other than an end node.

For each node type, the possible options are detailed below.

Decision, chance and logical nodes :



The left part, with the different types of nodes, updates the right part according to the selected type of node. The decision, chance and logic nodes have similar settings because they are all nodes with branches. So we will present them together.

- **Name**: Enter the node name.
- **Branches** :

  This is where the branches of the node are defined. You must enter different parameters for each branch:

  - **Name**: The branch name.
  - **Force**: This option allows you to force a branch to be part of the optimal path. Only one branch can be forced into the whole tree.
  - **Value**: The cost or profit associated with the branch. If it is a cost then the value must be negative. This parameter can refer to the value of a cell on a sheet. In this case, select the relevant cell.
  - **Probability**: The probability of completion associated with the branch. This parameter is only available for a chance node. This parameter can refer to the value of a cell on a sheet. In this case, select the relevant cell.
  - **Logical**: The logical formula associated with the branch. This parameter is only available for a logical node. This parameter can refer to the value of a cell on a sheet. In this case, select the relevant cell. It is also possible to directly write a formula with a logical result (TRUE or FALSE).

  The **Value** and **Reference** options allow you to choose the display mode of data that can refer to a cell in a sheet (Value, Probability and Logical). The list of branches is automatically updated according to your choice. These options only apply to the dialog box and not to the information displayed in a branch block.

In order to modify a branch, all you have to do is select it in the list of branches, modify its parameters in the **Selection** part and update in the list of branches by clicking the ✓ button.

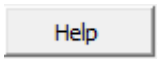
Other actions are possible at the level of the list of branches:

- ✚: Click this button to add a new branch in the list of branches. This will be added above the selected branch.
- ✖: click this button to remove the selected branch from the list of branches.
- ↑: Click this button to move the selected branch upwards in the list of branches.
- ↓: Click this button to move the selected branch down in the list of branches.

Once you have set all options:

: Click this button to validate your parameters and start building a new tree or updating the parameters of an existing tree.

: Click this button to close the dialog box without creating a new tree or modifying the parameters of an existing tree.



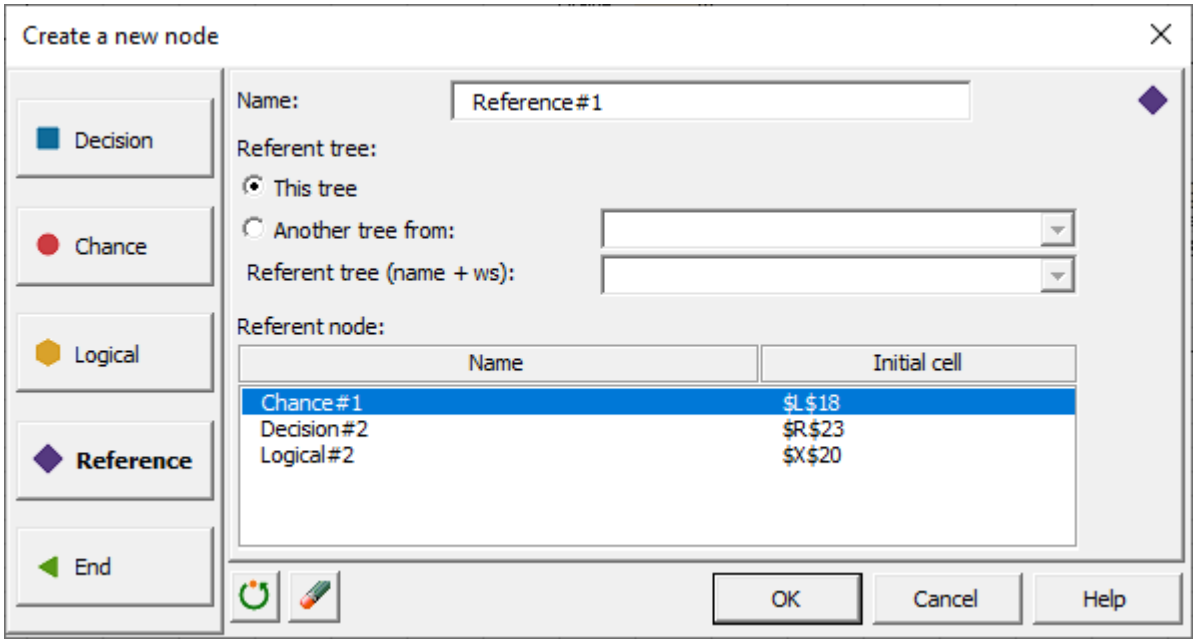
: Click on this button to display the help file relating to decision trees in XLSTAT.


: Click this button to restore the default options.


: Click this button to clear data selections.


: Click the orange button to save the dialog box settings to a file, or the green button to load the dialog box settings from a file.

Reference node:



.

- **Name**: Enter the node name.
- **Referent tree** :
  - **This tree**: Select this option if the referent tree is the same as the active tree.
  - **Another tree from**: Select this option if the referent tree is not the active tree. Then select, in the adjacent list, the workbook in which the referent tree is located.
  - **Referring tree (name + ws)**: Select the referent tree from this list. The list is made up of all the trees present in the previously selected workbook. Only the name of the tree is displayed, but when the list is expanded, the name of the leaf where the tree is located is also entered for more details.
- **Referent node**:
  - **Name**: Name of all the nodes present in the tree selected in the previous section.
  - **Initial cell**: Initial cell of the node block.

  Select the referent node from the list. If it is in the active workbook, then the corresponding block is selected in the sheet where it is located. This allows you to find your way around better if necessary.

Once the options have been have set:

OK: Click this button to validate your settings and start building a new tree or update the settings of an existing tree.

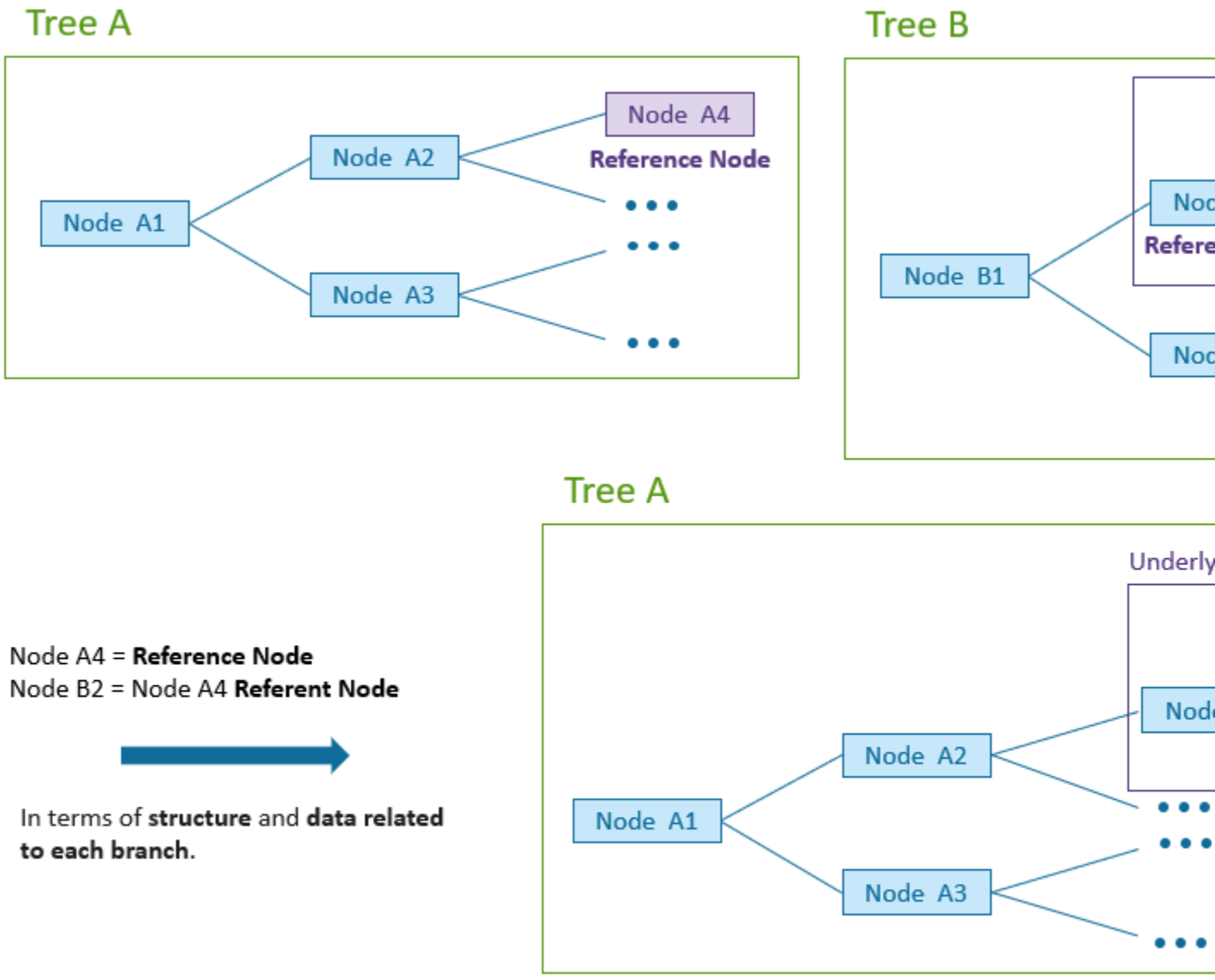
Cancel: Click this button to close the dialog box without creating a new tree or modifying the parameters of an existing tree.

Help: Click on this button to display the XLSTAT decision trees help file.

: Click this button to restore the default options.

: Click this button to clear data selections.

Here is a diagram to help you understand the notions of reference node and referent node :



Node A4 of the tree A inherits the following properties from node B2 of the tree B :

- Child nodes
- Child branches : values (costs and profits), probabilities (if chance node), logical formulas (if logical node)

Warning: parent nodes and branches remain unchanged.

Note : If a tree A contains a reference node whose referent node is in tree B, then tree B cannot have a referent node in tree A. You must have the two reference nodes in tree A.

Actions on a tree

Directly in blocks that constitute the tree

You can modify some information relating to a tree directly in the Excel sheet where it is located. The affected cells are colored. Here is the list of things you can change:

- Tree name
- Node name
- Branch name
- Branch value
- Branch probability (in the case of a chance node)
- Logical formula of a branch (in the case of a logical node)

Click on a node icon

- Decision:
- Chance:
- Logical:
- Reference:
- End:

If you click the icon of a node, the dialog box allowing you to configure it opens. You can then modify its settings (even its type) and validate it by clicking the *Continue* button.

Right click on a tree block

If you right click on the tree block you will be able to access an XLDTREE menu:

Here is the list of the different possible actions:

- **Open the dialog box to define settings for the selected tree**: Click on this if you want to display the tree settings dialog box.
- **Highlight the optimal path for the selected tree**: Click this item if you want to highlight the optimal path for the entire tree. The notion of optimal path is discussed in the Calculations and optimal path section.

- **Remove the optimal path for the selected tree**: Click on this if you no longer want to highlight the optimal path for the entire tree.
- **Delete the selected tree**: click on this if you want to safely delete the tree.

Right click on a node block

If you right click on a node block you will be able to access an XLDTREE menu:

Here are the possible actions:

- **Create a new node**: Click on this icon, available in an end node, if you want to display the dialog box for configuring a node and thus create a new node in place of the selected end node.
- **Open the dialog box to define settings for the selected node**: Click on this if you want to display the dialog box for configuring a node in order to update the selected node.
- **Highlight the optimal path from the selected node**: click this if you want to highlight the optimal path from this node. The notion of optimal path is discussed in the Calculations and optimal path section.
- **Remove the optimal path starting at the selected node**: click on this if you no longer want to highlight the optimal path starting from this node.
- **Insert a new node before the selected node**: Click on this if you want to insert a new node before the target node. By default, a decision node with two branches will be inserted. The target node will be on the first branch of this new node.
- **Delete the subtree from the selected node**: Click on this if you want to delete the target node as well as all of its child nodes. The target node will be replaced by an end node.
- **Copy the subtree from the selected node**: Click on this if you want to copy the subtree composed of the target node with all of its child nodes.
- **Paste the sub-tree in place of the selected node and its children**: Click on this if you want to replace the sub-tree made up of the target node and its child nodes by the sub-tree made up of the previously copied node with its child nodes.

Right click on a branch block

Right click on a branch block to access an XLDTREE menu:

The possible actions are:

- **Add a new branch above the selected branch**: Click on this icon if you want to add a new branch. It will be added above the target branch.
- **Delete the selected branch and its children**: click on this if you want to delete the target branch with all of its children.
- **Move the selected branch upwards**: Click on this if you want to move the target branch upwards (in the case of a tree displayed horizontally).
- **Move the selected branch down**: Click on this if you want to move the target branch down (in the case of a tree displayed horizontally).
- **Move the selected branch to the left**: click on this if you want to move the target branch to the left (in the case of a tree displayed vertically).
- **Move the selected branch to the right**: Click on this if you want to move the target branch to the right (in the case of a tree displayed vertically).
- **Force the selected branch to be on the optimal path** : click on this if you want to force the optimal path to go through the target branch. This can be useful when you return to an existing tree later, when certain choices have already been made or carried out. A single branch, among all those of the tree, can be forced. If another branch was already forced then it will no longer be. The forced branch has its name displayed in a different color than the other branch names. The notion of optimal path is discussed in the section Calculations and optimal path.
- **Stop forcing the selected branch to be on the optimal path** : click on this if you want to stop forcing the optimal path going through the target branch. No more branches will be forced in the active tree.

Calculations and optimal path

Calculations

For each block of a tree, whether it is the block of the tree itself, a node block or a branch block, calculations are carried out for each modification. Their result is displayed directly in the form of formulas, in the blocks concerned. Unlike editable information where cells are colored, the information resulting from the calculation of a formula is on a white background. You can replace formulas deleted by mistake by generating the tree again. All you have to do is open the tree parameterization dialog box and click on the OK button..

As seen previously, it is possible to choose which information is displayed for the node blocks in the tree settings dialog box.

Optimal path

This tool is not yet available on the Mac platform.

The optimal path represents the path that best meets the chosen calculation mode. It depends on the expected gain, with utility or not, from each node. The choice of an optimal path for the expected gain or the expected gain with utility is done via the tree parameterization dialog box. The behavior is different depending on the type of node:

- **Decision node**: The optimal path goes through the branch whose gain, with utility or not, best meets the chosen calculation mode.
- **Chance node**: The optimal path arriving at a chance node passes through all the branches of this node. Indeed, it is not possible to know in advance which branch will be carried out, no matter how likely it is. The gain, with utility or not, takes this behavior into account since it is the weighted sum (by the probability of realization) of the gain, with utility or not, expected from the child node of each branch.
- **Logical node**: The optimal path goes through the only branch with a TRUE result. As already seen, it is not possible to have several branches with a TRUE result. If this is the case then the gain, with utility or not, and therefore the optimal path, cannot be calculated and is in error. If all the branches have a FALSE result, then the gain, with utility or not — and therefore the optimal path — are not available for this node.
- **Reference node**: A reference node has a decision, chance or logic node as its referent. It is therefore sufficient to refer to these types of nodes to know the behavior of the optimal path at the level of this type of node. However, this behavior will remain invisible because the reference node of a tree does not display its child nodes.

The optimal path, once activated, is highlighted on the tree. It can concern the entire tree or it can only start from a node that you have chosen by right-clicking. As long as the optimal tree is activated, it is automatically recalculated and its display updated, each time the tree is modified.

Examples

Examples showing how to create and use a decision tree are available on the XLSTAT Help Center at:

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

http://www.xlstat.com/demo-dct2.htm

http://www.xlstat.com/demo-dct3.htm

Bayesian networks

The Bayesian Networks module of XLSTAT allows statistical analysis by means of a Bayesian network. Very popular in artificial intelligence it is mainly used to represent knowledge and its uncertainties. It is a decision-making tool which main function is to show causal relationships between variables.

Bayesian networks are used in the finance, for example to analyze risks of credit card fraud, in medicine for example to make a diagnosis, or in industrial applications.

In this section :

Description

A Bayesian network is a statistical tool that allows to model dependence or conditional independence relationships between random variables. This method emerged from Judea Pearl's pioneering research in 1988 on the development of artificial intelligence techniques.

The originality of this method is that it offers a formal framework to represent the relational structure of network variables. On one side we have a qualitative description with a graph and on the other side a quantitative description with probabilities. The graph is a way to schematize the network with nodes and arcs connecting these nodes, which will facilitate the understanding of the problem and the interpretation of results.

In a Bayesian network, the graph must be oriented and must respect the so-called "acyclic" rule that is, it does not have a circuit (see Figure 1). In this case, one talks about an directed acyclic graph (DAG). When an arc starts from a node A and goes to a node B, we say that node A is a parent of node B and that node B is a child of node A. The probabilities are given for each node, according to the status of the latter in the network (child or parent).

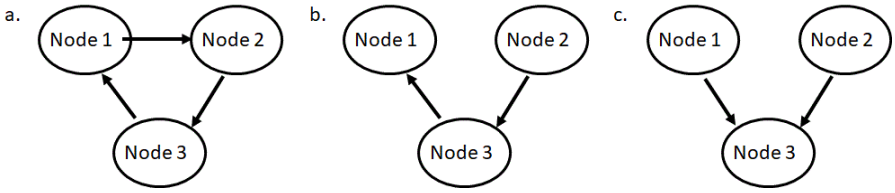


Figure 1. Example of oriented graphs with 3 nodes. Graph a. is a circuit while graphs b. and c. are acyclic.

A Bayesian network is used to make queries that is to evaluate the model and obtain probabilities a posteriori given new information. Let U be the set of nodes making up the network and $P(U)$ be the probability distribution on that set. If we have new information, denoted $epsilon$, on one or more variables, then we would like to update the knowledge of the Bayesian network through $P(U)$ for this new information. This update is called inference. We can dissociate 2 types:

- from effect to cause: the new information comes from a child node and spreads to its parents. In this situation the Bayesian network serves to establish a diagnosis.
- from cause to effect: the new information comes from a parent node and spreads to its children. In this situation the Bayesian network is used to make simulations or predictions.

From a mathematical point of view, the inference in a Bayesian network is the computation of $P(U|\epsilon)$, that is the computation of the a posteriori probability of the network knowing $ epsilon$. Two equations are fundamental in this computation: the Bayes theorem which is written according to the following equation:

$$P(A|B) = \frac{P(A|B)P(A)}{P(B)},$$

where A and B are two random variables, and where the joint probability formula is given by:

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A).$$

The term $P(A|B)$ is called the conditional probability of A knowing B (or A conditional B). The term $P(B|A)$, for a known A, is called the likelihood function of B. The terms $P(A)$ and $P(B)$ are respectively the a priori probability of A or the marginal probability of A and the marginal or a priori probability of B.

However this computation can be complex. Depending on the complexity of the network (the simpler the network topology, the easier the inference), the computation of the joint probability, for example, which is the product of conditional and marginal probabilities for all the values of the network variables, becomes time-consuming. It is even proven that in the general case this is a NP-difficult problem (Cooper G. F., 1990).

From this point were born several algorithms allowing the computation in a complex probabilistic system. XLSTAT uses the exact inference algorithm known as the junction tree (Jensen et al., 1990). Its principle is to transform the DAG into a tree and then to derive its junction tree. The latter is also a graph whose node is called a clique. In graph theory, a clique is a subset of fully connected nodes of the original graph. In the example given in Figure 2, the graph b is formed by 4 cliques all composed of 2 nodes of the original graph a. The joined probability of the network thus becomes the product of the joint probabilities of each clique.

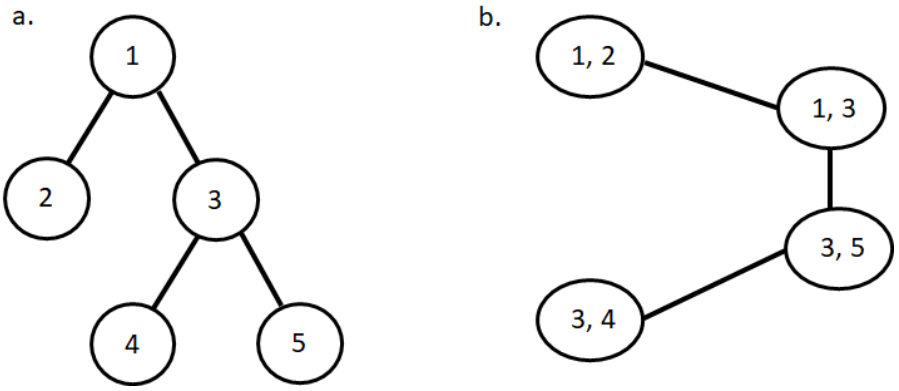


Figure 2. Tree (a.) and junction tree (b.) examples.

The procedure for analysing a Bayesian network in XLSTAT is as follows:

1. Open a project: in the XLSTAT menu go to the Bayesian Networks module and open a new project. A window opens offering the choice between the classic and expert mode (see paragraph Options and object selection on the graph for more information). In both cases a workbook will open consisting of several sheets as described in Projects that will be used to create your network.

2. Construct the graph: draw your network using the provided buttons and name each of your nodes. For more information on this step go directly to Graph construction.

3. Definition of probability distributions: for each variable you must fill in a probability table. This can be done in two ways: expert mode, where the network is built from human knowledge, or learning mode, where the network can automatically be learned from data. These two options are described in the section Probability tables definition.

4. Inference: You can now query your Bayesian network and run probability calculations from the dedicated button. For more information on using this feature read the section Bayesian network analysis.

Projects

Bayesian Networks projects are specific Excel workbooks. When you create a new project, its default name starts with BNbook. You can save it under a name of your choice. The "Save" button in the options provided in the Bayesian Networks module allows you to save your project using the extension *.xlsm.

A raw Bayesian Networks project always initially contains two sheets and then three when the probability tables are defined. These sheets, which should not be deleted, are:

- Data : it is an empty Excel sheet in which your data must be copied/pasted.
- BNGraph : it is a sheet containing a drawing area, blank at the beginning, and a toolbar whose description is made in the next section. It must be used to draw the graph of the Bayesian network.
- Probability tables: this sheet contains the probability distributions of all nodes that were drawn on the BNgraph sheet. They are formalized in the form of one table for each node. For each table the number of columns is the number of parents plus one column for the node itself, plus one column for the probabilities. A table has at least two columns. For maginal nodes, there are only two columns (as it has no parent). The number of rows is defined by the number of modality combinations of the node and its parents. For example, there are 3 nodes A, B and C with 2, 3 and 2 modalities respectively whose labels are $\{a_1, a_2\}$, $\{b_1, b_2, b_3\}$, $\{c_1, c_2\}$. Nodes B and C are the parents of node A. The probability table of node A takes the form of:

A	B	C	Probability
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	----------

The order of the parents is decreasing with the number of modalities. The order of modalities in the table is the order given by the user.

Once the Bayesian network is drawn and the probability tables are filled, you can start the probability computations. The results are displayed in an Excel sheet following the BNGraph sheet.

It is possible to save a model before modifying it in order to be able to modify it later (see section Options and object selection on the graph for more details).

Toolbars

A toolbar is available at the top left of the BNGraph sheet which includes huit orange buttons aligned horizontally:

These buttons are useful in the construction of the graph, in the definition of probability tables and in the analysis of the Bayesian network. More precisely the first four buttons are dedicated to the drawing of the Bayesian network, the next two to fill in the probability table, the seventh is used to compute the probabilities of the Bayesian network drawn and the last button redirects to this help. The mode of use and functionality of all these buttons are described in the following sections of this help.

The toolbar is only visible when you are on the BNGraph sheet.

Options and object selection on the graph

The Bayesian Networks module of XLSTAT offers 3 options:

 Click this button to open a new Bayesian Networks project. A window then opens offering you the choice between a method display in classic or expert mode. Classic mode is preferred when you have a data set and you want the probability distributions to be computed automatically. The expert mode is intended as its name indicates to "experts" of the problem studied because it allows the user to define himself the modalities and probabilities of each variable. In both cases a new project file will open, for which you can find a description in the previous section.

 Click this button to open an existing Bayesian Network project.

 Click this button to save the active Bayesian Network project. This button is only accessible if changes have been made in the project.

The selection of objects (node or arc) on the BNGraph sheet is done only with the Ctrl key + left click with the mouse, or cmd + left click for Mac users. Do not use the right mouse click and the Excel drop-down menu.

Deselection is done with the Esc key and applies to all selected objects. You can move the objects with the mouse, making sure you have deselected everything before and then selecting the desired objects. You can also use the keyboard arrows. You can delete a node or an arc, only one at a time, by selecting it and pressing the Delete key on your keyboard. When you connect two nodes by an arc, this later should not be moved to connect two other nodes as it was defined for the first two nodes only.

Below the button bar, greyed-out text summarizes the available keyboard shortcuts, which are:

- Node or arc selection: Ctrl + left-click (or cmd + left-click if you are a Mac user)
- Unselect all: Escape
- Move node: Deselect all + select node + move with mouse
- Delete: Ctrl + Delete
- Link 2 nodes: Ctrl + A
- Change arc direction: Ctrl + D
- Modality Editeur: Ctrl + E (visible in expert mode only)
- Data Editeur: Ctrl + L (visible in classic mode only)

Graph construction

The graph of a Bayesian network is materialized by nodes and arcs that link nodes. In a Bayesian networks project (see projects), the graph must be drawn on the sheet named BNGraph using the toolbar. The selection of nodes and arcs is done according to the description made in the Options and object selection on the graph section.

Click this button to add a node on BNGraph sheet. Once the button activated it becomes grey meaning that you are allowed to draw the node on the sheet where you want by clicking on it. After that, a window opens to name the node. You can rename it later or change its name by simply clicking on the node again.

. This button is used to name the nodes. To do this select first a node and then click on this button. You can name only one node at a time.

Click this button to link two nodes with an arc. To use it you must first select the parent then the child and then the Arc button. You cannot link more than two nodes at a time and you can create only one arc between the same two nodes.

Click this button to change the direction of an arc between two nodes. To use it you must first select the arc of your choice and click the button. You cannot change the direction of several arcs at a time.

Probability tables definition

The variables involved in the Bayesian network can be qualitative or quantitative. They have at least 2 modalities. The probability tables indicate the different values taken for these modalities. When a variable is dependent on one or more variables the table of probabilities gives the values taken for all modality combinations of the set of variables.

In a Bayesian Networks project you can define the probability tables in two ways:

- in classic mode,
- in expert mode.

The use of these two modes is explained in the following two sections. In both cases the probability tables are displayed in a dedicated Excel sheet as described in the projects section. The button  , more useful in expert mode, is common to both modes. It allows to view and/or modify the value of a specific probability. A detailed description of its use can be found in the expert section.

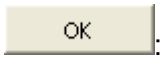
Classic mode

Modalities and probability tables are automatically computed from a dataset. The column label must contain the variable name. Be careful to make sure that their name is the same as the one used to the nodes in your network.

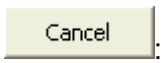
The algorithm works with qualitative variables. When quantitative variables are selected, they are transformed into qualitative ones as follows. First, for each of them, their values are sorted in ascending order, then discretized in a maximum number of ten intervals. These new intervals are then used to recode the values and become the modalities. A table of modalities is then calculated for all variables taking into account their relational structure in the network. This table is similar to the one presented in the paragraph projects without the probability column. The frequency of occurrence of these modalities in the data is then computed, and then converted into probability under the condition of marginal sum equals to 1(the sum of the values taken by a variable for a given value of the other variables). When the frequency is zero for more than half of the variable's modalities, the probabilities are not computed by the algorithm but replaced by a missing value in the result sheet. Conversely when the frequency is zero for less than half then the probabilities are computed as the mean of the difference of 1 and the probabilities already calculated.

To use this feature click the following button to load all your data:

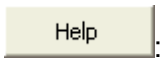
This action results in the display of a dialog box composed of several tabs corresponding to the different options available both for the management of computations and for the display of results. Below you will find a description of the different elements of the dialog.



: click this button to start the computations.



: click this button to close the dialog box without performing the computations



: click this button to display the help.

 : click this button to restore the default options.

 : click this button to clear data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Qualitatives data: activate this option if you want to use qualitative variables and then select these variables.

Quantitatives data: activate this option if you want to use quantitative variables and then select these variables.

Observation weights: activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice.

Missing data tab:

Do not accept missing data: activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: activate this option to ignore the observations that contain missing data.

Estimate missing data: activate this option to estimate the missing data before the calculation starts.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data for an criterion by searching for the nearest neighbor of it.

Results tab:

The results are displayed in a new sheet called probability tables. A summary of the data is first presented at the beginning of the sheet specifying the number of missing data, the replacement value if this option was chosen and the descriptive statistics of the variables. Next comes the probability distribution of each variable according to the status of the latter in the Bayesian network. These results are tables like the one presented in the projects section. At the end of the sheet you will find a button to directly launch the analyses on the Bayesian network evaluated with these probability tables.

In expert mode

It is the user who defines, step by step, all the required information on the variables, namely the modalities and the probabilities. For this purpose two buttons are available to you for one and the other action.

 Click this button to set the modalities of your variables. A window opens allowing you to select data from the Data sheet (see the projects section). In this sheet are listed in column the modalities of the variables with one column per variable. The column header contains the name of the variables and must match the name of the nodes drawn on the BNGraph sheet. You can select one or more columns. If you want to define the modalities of a single variable you can preselect the corresponding node on the BNGraph sheet, the name of the selected node will then appear in the modality editor window. When you click Ok the probability tables of the selected variables are updated in the dedicated sheet.

 Click this button to display and/or modify the probabilities of a variable. To do this you must first select a node on the BNGraph sheet (see the section Options and object selection on the graph for the node selection). A window then opens and displays the probability table of the selected variable. You can change the value of one probability at a time. To do this, you need to select it and click the "edit" button. A window opens in which you can enter a new value. When you click OK, the value updated in the probability table. If you click OK again this value will be saved in the excel sheet where the probability tables for all variables are saved.

Analysis of a Bayesian network

Once the model is designed on the BNGraph sheet, and once all the probabilities have been defined for each variable, you can click the  button of the toolbar to launch the analyses on the Bayesian network. This is also possible from the button at the end of the probability tables sheet. A dialog box is displayed with the following tabs.

General tab:

Data source: select a sheet containing the XLSTAT generated probability tables (see Probability tables definition) of the Bayesian network you wish to analyze.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Missing data tab:

Do not accept missing data: activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Estimate missing data: activate this option to estimate the missing data before the calculation starts.

- **Mean**: activate this option to estimate the missing data by using the mean (quantitative variables) of actions.

Outputs tab:

Marginal probabilities: activate this option to display the distribution of the marginal probabilities of each node/variable.

Join probability distribution: activate this option to display the join probability distribution of each clique (see Description of the method for definition).

Conditional probabilities: activate this option to display the conditional probability distribution of each node/variable.

Graphs tab:

Marginal probabilities charts: activate this option to display the graph of the marginal probability distribution of each node/variable.

Conditional probabilities charts: activate this option to display the graph of the conditional probability distribution of each node/variable.

Results

The results obtained answer all possible queries on the Bayesian network analyzed.

Marginal probability distribution of each node: this result, given as a table, corresponds to the marginal probabilities of each node drawn on the BNGraph sheet. If the graph option is selected the result is also displayed as a bar chart under the table.

Join probability distribution for each clique: This result displays the number of cliques, the list of nodes involved in each of cliques and their correlated join probability distribution table.

Conditional probability distribution of each node: this result,given as a table, corresponds to the conditional probabilities of each dependent node. If the graph option is selected the result is also displayed as a bar chart under each table.

Example

A tutorial on Bayesian network tool is available on the XLSTAT Help Center:

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

References

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.

Time series analysis

Time series visualization

Use this tool to create in three clicks as many charts as you have time series.

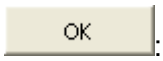
In this section:

Description

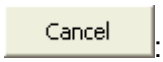
This tool allows to create I three clicks as many charts as you have time series. It also allows you to group the series on a single graph. Finally, an option allows you to link charts to the input data: If you choose that option, charts are automatically updated when there is a change in the input data.

Dialog box

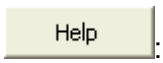
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Times series: Select the data that correspond to the time series. If a header is available on the first row, make sure you activate the "Series labels" option.

Date data: Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format. If this option is not activated, XLSTAT creates its own time variable ranging from 1 to the number of data.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Series labels: Activate this option if the first row of the selected series includes a header.

Charts tab:

Link the chart to the input data: Activate this option so that a change in the input data directly results in an update of the chart.

Display all series on a single chart: Activate this option to display the data on a single chart.

Results

Charts are displayed for all the selected series.

Example

An example of time series visualization is available at the XLSTAT Help Center:

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

References

Brockwell P.J. and Davis R.A. (1996). Introduction to Time Series and Forecasting. Springer Verlag, New York.

Descriptive analysis

Use this tool to compute the descriptive statistics that are specially suited for time series analysis.

In this section:

Description

One of the key issues in time series analysis is to determine whether the value we observe at time t depends on what has been observed in the past or not. If the answer is yes, then the next question is how.

The sample autocovariance function (ACVF) and the autocorrelation function (ACF) give an idea of the degree of dependence between the values of a time series. The visualization of the ACF or of the partial autocorrelation function (PACF) helps to identify the suitable models to explain the passed observations and to do predictions. The theory shows that the PACF function of an AR(p) – an autoregressive process of order p - is zero for lags greater than p.

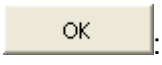
The cross-correlations function (CCF) allows to relate two time series, and to determine if they co-vary and to which extent.

The ACVF, the ACF, the PACF and CCF are computed by this tool.

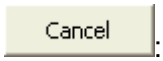
One important step in time series analysis is the transformation of time series (see Transforming time series) which goal is to obtain a white noise. Obtaining a white noise means that all deterministic and autocorrelations components have been removed. Several white noise tests, based on the ACF, are available to test whether a time series can be assumed to be a white noise or not.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help options.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Times series: Select the data that correspond to the time series. If a header is available on the first row, make sure you activate the "Series labels" option.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Series labels: Activate this option if the first row of the selected series includes a header.

Options tab:

Time steps: the number of time steps for which the statistics are computed can be automatically determined by XLSTAT, or set by the user.

Confidence interval (%): Activate this option to display the confidence intervals. The value you enter (between 0.01 and 99.99) is used to determine the confidence intervals for the estimated values. Confidence intervals are automatically displayed on the charts.

- **White noise assumption**: Activate this option if you want that the confidence intervals are computed using the assumption that the time series is a white noise.

White noise tests: Activate this option if you want XLSTAT to display the results of the normality test and the white noise tests.

- **h1**: Enter the minimum number of lags to compute the white noise tests.
- **h2**: Enter the maximum number of lags to compute the white noise tests.
- **s**: Enter the number of lags between two series of white noise tests. s must be a multiple of (h2-h1).

Significance level (%): Enter the significance level for the tests (default value: 5%).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Replace by the average of the previous and next values: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

Ignore missing data: Activate this option to ignore missing data.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics of the selected series.

Autocorrelations: Activate this option to estimate the autocorrelation function of the selected series (ACF).

Autocovariances: Activate this option to estimate the autocovariance function of the selected series.

Partial autocorrelations: Activate this option to compute the partial autocorrelations of the selected series (PACF).

Cross-correlations: Activate this option to compute the estimate of the cross-correlation function (CCF).

Charts tab:

Autocorrelogram: Activate this option to display the autocorrelogram of the selected series.

Partial autocorrelogram: Activate this option to display the partial autocorrelogram of the selected series.

Cross-correlations: Activate this option to display the cross- correlations diagram in the case where several series have been selected.

Results

For each series, the following results are displayed:

Summary statistics: This table displays for the selected variables, the number of observations, the number of missing values, the number of non- missing values, the mean and the standard deviation (unbiased).

Normality and white noise tests: Table displaying the results of the various tests. The Jarque-Bera normality test is computed once on the time series, while the other tests (Box-Pierce, Ljung-Box and McLeod-Li) are computed at each selected lag. The degrees of freedom (DF), the value of the statistics and the p-value computed using a Chi-Square(DF) distribution are displayed. For the Jarque-Bera test, the lower the p-value, the more likely the normality of the sample. For the three other tests, the lower the p-value, the less likely the randomness of the data.

Descriptive functions for the series: Table displaying for each time lag the values of the various selected descriptive functions, and the corresponding confidence intervals.

Charts: For each selected function, a chart is displayed if the "Charts" option has been activated in the dialog box.

If several time series have been selected and if the "cross-correlations" option has been selected the following results are displayed:

Normality and white noise tests: Table displaying the results of the various tests, Box-Pierce, Ljung-Box and McLeod-Li, which are computed at each selected lag. The degrees of freedom (DF), the value of the statistics and the p-value computed using a Chi-Square(DF) distribution are displayed. The lower the p-value, the less likely the randomness of the data.

Cross-correlations: Table displaying for each time lag the value of the cross-correlation function.

Example

A tutorial explaining how to use descriptive analysis with a time series is available on the XLSTAT Help Center. To consult the tutorial, please go to:

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

References

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.

Mann-Kendall Trend Tests

Use this tool to determine with a nonparametric test if a trend can be identified in a series, even if there is a seasonal component in the series.

In this section:

Description

A nonparametric trend test was first proposed by Mann (1945), then further studied by Kendall (1975) and improved by Hirsch *et al* (1982, 1984) who have made it possible to take seasonality into account.

The null hypothesis H_0 for these tests is that there is no trend in the series. The three alternative hypotheses that there is a negative, non-null, or positive trend can be chosen.

The Mann-Kendall tests are based on the calculation of Kendall's tau measure of association between two samples, which is itself based on the ranks with the samples.

Mann-Kendall trend test

In the particular case of the trend test, the first series is an increasing time indicator generated automatically for which ranks are obvious, which simplifies the calculations. The S statistic used for the test and its variance are given by:

$$S = \sum_{i=1}^{x-1} \sum_{j=i+1}^{x} \operatorname{sgn}(x_j - x_i)$$

$$Var(S) = \frac{n(n-1)(2n+5)}{18}$$

where n is the number of observations and x_i $(i = 1 \ldots n)$ are the independent observations.

To calculate the p-value of this test, XLSTAT can calculate, as in the case of the Kendall tau test, an exact p-value if there are no ties in the series and if the sample size is less than 50. If an exact calculation is not possible, a normal approximation is used, for which a correction for continuity is optional but recommended.

Considering the autocorrelations

The Mann-Kendall trend test requires that the observations are independent (meaning the correlation between the series with itself with a given lag should not be significant). In the case where there is some autocorrelation in the series, the variance of the S statistic has been shown to be underestimated.

Therefore, several improvements have been suggested. XLSTAT offers two alternative methods, the first one published by Hamed and Rao (1998) and the second by Yue and Wang (2004). Both methods imply the computation of Sen's slope estimator (Sen, P. K. (1968)). The value of which is displayed by XLSTAT if the corresponding option is activated in the output tab.

Before running a Mann-Kendall trend test, it is advisable to check the autocorrelations of the series under study using XLSTAT's descriptive statistics.

Hamed and Rao's modified variance

The first method performs well in the case of no trend in the series, it avoids identifying a trend when it is in fact due to the autocorrelation. The variance depends on the autocorrelations that have been estimated. The formula is as follows (Hamed and Rao, 1998):

$$Var(S)_{HamedRao} = Var(S) \cdot (1 + \frac{2}{n(n-1)(n-2)} \times \sum_{i=1}^{n-1}(n-i)(n-i-1)(n-i-2)\rho_s(i)),$$

where $\rho_s(i)$ is the autocorrelation function of the ranks of the observations.

Note: As Hamed and Rao (1998) put it, positive autocorrelation leads to an increase in V(S). Negative autocorrelation in the data has the opposite effect, reducing the variance of S. The Hamed and Rao corrected variance calculation can in fact show a negative variance if the autocorrelation function has negative values.

Yue and Wang's modified variance

This second method has the advantage of performing better when there are both a trend and an autocorrelation.

$$Var(S)_{HamedRao} = Var(S) \cdot (1 + \frac{2}{n} \cdot \sum_{i=1}^{n-1}(n-i)\rho(i)),$$

where $\rho(i)$ is the lag-i serial correlation coefficient.

Seasonal Mann-Kendall test

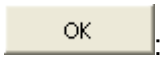
In the case of seasonal Mann-Kendall test, we take into account the seasonality of the series. This means that for monthly data with seasonality of 12 months, one will not try to find out if there is a trend in the overall series, but if from one month of January to another, and from one month of February and another, and so on, there is a trend.

For this test, we first calculate all Kendall's tau for each season, then calculate an average Kendall's tau. The variance of the statistic can be calculated assuming that the series are independent (e.g. values of January and February are independent) or dependent, which requires the calculation of a covariance. XLSTAT allows both (serial dependence or not).

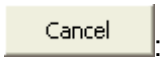
To calculate the p-value of this test, XLSTAT uses a normal approximation to the distribution of the average Kendall tau. A continuity correction can be used.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Times series: Select the data that corresponds to the time series. If a header is available on the first row, make sure you activate the "Series labels" option.

Date data: Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Series labels: Activate this option if the first row of the selected series includes a header.

Mann-Kendall trend test: Activate this option to run this test.

Seasonal Mann-Kendall test: Activate this option to run this test. Then enter the value of the **period** (number of lags between two seasons). Specify if you consider that there is serial dependence or not.

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see the description section for more details).

Significance level (%): Enter the significance level for the test (default value: 5%).

Exact p-values: Activate this option if you want XLSTAT to calculate the exact p-value as far as possible (see description).

Continuity correction: Activate this option if you want XLSTAT to use the continuity correction if the exact p-values calculation has not been requested or is not possible (see description).

Autocorrelations: Activate one of the two options **Hamed and Rao** or **Yue and Wang** to into account for autocorrelations in the series. For the Hamed and Rao option you can filter out the autocorrelations for which the p-value is not below a given level that you can set (default value: 10%).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Replace by the average of the previous and next values: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

Ignore missing data: Activate this option to ignore missing data.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics of the selected series.

Sen's slope: Activate this option to display Sen's slope estimator. You can also configure the Confidence Interval.

Charts tab:

Display charts: Activate this option to display line plot of the data.

Results

For each series, the following results are displayed:

Summary statistics: This table displays for the selected variables, the number of observations, the number of missing values, the number of non- missing values, the mean and the standard deviation (unbiased).

Mann-Kendall trend test: Results of the Mann-Kendall trend test are displayed if the corresponding option has been activated. It is followed by an interpretation of the results.

Seasonal Mann-Kendall trend test: Results of the seasonal Mann- Kendall test are displayed if the corresponding option has been activated. It is followed by an interpretation of the results.

Sen's slope: The value of the Sen's slope is given. The closer it is to 0, the lesser the trend. The sign of the slope tells if the trend is increasing or decreasing.

Example

A tutorial explaining how to use the Mann-Kendall trend tests with a time series is available on the XLSTAT Help Center. To consult the tutorial, please go to:

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

References

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.

Homogeneity tests

Use this tool to determine using one of four proposed tests (Pettitt, Buishand, SNHT, or von Neumann), if we may consider a series is homogeneous over time, or if there is a time at which a change occurs.

In this section:

Description

Homogeneity tests involve a large number of tests for which the null hypothesis is that a time series is homogenous between two given times.

The variety of the tests comes from the fact that there are many possible alternative hypotheses: change in distribution, changes in average (one or more times) or presence of trend.

The tests presented in this tool correspond to the alternative hypothesis of a single shift. For all tests, XLSTAT provides p-values using Monte Carlo resamplings. Exact calculations are either impossible or too costly in computing time.

When presenting the various tests, by $X_i(i = 1, 2, \ldots, T)$ we refer to a series of T variables for which we observe $x_i(i = 1, 2, 3, \ldots, T)$ at T successive times. Let $\hat{\mu}$ be the mean of the T observed values and let $\hat{\sigma}$ be the biased estimator of their standard deviation (we divide by T).

Note 1: If you have a clear idea of the time when the shift occurs, one can use the tests available in the parametric or nonparametric tests sections. For example, assuming that the variables follow normal distributions, one can use the test z (known variance) or the Student t test (estimated variance) to test the presence of a change at time t. If one believes that the variance changes, you can use a comparison test of variances (F-test in the normal case, for example, or Kolmogorov-Smirnov in a more general case).

Note 2: The tests presented below are sensitive to a trend (for example a linear trend). Before applying these tests, you need to be sure you want to identify a time at which there is a shift between two homogeneous series.

Pettitt's test

The Pettitt's test is a nonparametric test that requires no assumption about the distribution of data. The Pettitt's test is an adaptation of the tank- based Mann-Whitney test that allows identifying the time at which the shift occurs.

In his article of 1979 Pettitt describes the null hypothesis as being that the T variables follow the same distribution F, and the alternative hypothesis as being that at a time τ there is a change of distribution. Nevertheless, the Pettitt test does not detect a change in distribution if there is no change of location. For example, if before the time τ, the variables follow a normal $N(0,1)$ distribution and from time τ a $N(0,3)$ distribution, the Pettitt test will not detect a change in the same way a Mann-Whitney would not detect a change of position in such a case. In this case, one should use a Kolmogorov Smirnov based test or another method able to detect a change in another characteristic than the location. We thus reformulate the null and alternative hypotheses:

H0: The T variables follow one or more distributions that have the same location parameter.

Two-tailed test: Ha: There exists a time τ from which the variables change of location parameter.

Left-tailed test: Ha: There exists a time τ from which the variables location is reduced by Δ.

right-tailed test: Ha: There exists a time τ from which the variables location is augmented by Δ.

The statistic used for the Pettitt's test is computed as follows:

Let $D_{ij} = -1$ if $(x_i - x_j) < 0$, $D_{ij} = 0$ if $(x_i - x_j) = 0$, $D_{ij} = 1$ if $(x_i - x_j) > 0$

We then define $U_{t,T} = \sum_{i=1}^{t} \sum_{j=i+1}^{T} D_{ij}$

The Petitt's statistic for the various alternative hypotheses is given by:

$K_T = \max_{1 \leq t < T} |U_{t,T}|$, for the two-tailed case

$K_T^+ = \max_{1 \leq t < T} U_{t,T}$, for the left-tailed case

$K_T^- = -\min_{1 \leq t < T} U_{t,T}$, for the right-tailed case

XLSTAT evaluates the p-value and an interval around the p-value by using a Monte Carlo method.

Alexandersson's SNHT test

The SNHT test (*Standard Normal Homogeneity Test*) was developed by Alexandersson (1986) to detect a change in a series of rainfall data. The test is applied to a series of ratios that compare the observations of a measuring station with the average of several stations. The ratios are then standardized. The series of X_i corresponds here to the standardized ratios. The null and alternative hypotheses are determined by:

H0: The T variables X_i follow a $N(0,1)$ distribution.

Ha: Between times 1 and ν the variables follow an $N(\mu_1,1)$ distribution, and between $\nu+1$ and T they follow an $N(\mu_2,1)$ distribution.

The Alexandersson statistic is defined by:

$$T_0 = \max_{1\leq t<T}[\nu\overline{z}_1^2 + (n-\nu)\overline{z}_2^2]$$

with

$$\begin{aligned}\overline{z}_1 &= \frac{1}{\nu}\sum_{t=1}^{\nu} x_t \\ \overline{z}_2 &= \frac{1}{n-\nu}\sum_{t=\nu+1}^{T} x_i\end{aligned}$$

The To statistic derives from a calculation comparing the likelihood of the two alternative models. The model corresponding to Ha implies that μ_1 and μ_2 are estimated while determining the ν parameter maximizing the likelihood.

XLSTAT evaluates the p-value and an interval around the p-value by using a Monte Carlo method.

Note: if ν is known, it is enough to run a z test on the two series of ratios. The SNHT test allows identifying the most likely ν.

Buishand's test

The Buishand's test (1982) can be used on variables following any type of distribution. But its properties have been particularly studied for the normal case. In his article, Buishand focuses on the case of the two-tailed test, but for the Q statistic presented below the one-sided cases are also possible. Buishand has developed a second statistic R, for which only a bilateral hypothesis is possible.

In the case of the Q statistic, the null and alternative hypotheses are given by:

H0: The T variables follow one or more distributions that have the same mean.

Two-tailed test: Ha: There exists a time τ from which the variables change of mean.

Left-tailed test: Ha: There exists a time τ from which the variables mean is reduced by Δ.

right-tailed test: Ha: There exists a time τ from which the variables mean is augmented by Δ.

We define $S_o^* = 0$, $S_k^* = \sum_{i=1}^{k}(x_i - \hat{\mu})$,$k = 1, 2, \ldots, T$ and $S_k^{**} = \frac{S_k^*}{\hat{\sigma}}$

The Buishand's Q statistics are computed as follows:

$Q = \max_{1\leq k<T}|S_k^{**}|$, for the two-tailed case

$Q^- = \max_{1 \leq k < T}(S_k^{**})$, for the left-tailed case

$Q^+ = -\min_{1 \leq k < T}(S_k^{**})$, for the right-tailed case

XLSTAT evaluates the p-value and an interval around the p-value by using a Monte Carlo method.

In the case of the R statistic (R stands for *Range*), the null and alternative hypotheses are given by:

- H0: The T variables follow one or more distributions that have the same mean.
- Two-sided test: Ha: The T variables are not homogeneous for what concerns their mean.

The Buishand's R statistic is computed as:

$$R = \max_{1 \leq k < T}(S_k^{**}) - \min_{1 \leq k < T}(S_k^{**})$$

XLSTAT evaluates the p-value and an interval around the p-value by using a Monte Carlo method.

Note: The R test does not allow detecting the time at which the change occurs.

von Neumann's ratio test

The von Neumann ratio is defined by:

$$N = \frac{1}{T\hat{\sigma}} \sum_{i=1}^{T-1} (x_i - x_{i+1})^2$$

We show that the expectation of N is 2 when the X_i have the same mean.

XLSTAT evaluates the p-value and an interval around the p-value by using a Monte Carlo method.

Note: This test does not allow detecting the time at which the change occurs.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

General tab:

Times series: Select the data that correspond to the time series. If a header is available on the first row, make sure you activate the "Series labels" option.

Date data: Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format. If this option is not activated, XLSTAT creates its own time variable ranging from 1 to the number of data.

- **Check intervals**: Activate this option so that XLSTAT checks that the spacing between the date data is regular.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Series labels: Activate this option if the first row of the selected series includes a header.

Pettitt 's test: Activate this option to run this test (see the description section for more details).

SNHT test: Activate this option to run this test (see the description section for more details).

Buishand's test: Activate this option to run this test (see the description section for more details).

von Neumann's test: Activate this option to run this test (see the description section for more details).

Options tab:

Alternative hypothesis: Choose the alternative hypothesis to be used for the test (see the description section for more details).

Significance level (%): Enter the significance level for the test (default value: 5%).

Monte Carlo method: Activate this option to compute the p-value using Monte Carlo simulations. Enter the maximum number of simulations to perform and the maximum computing time (in seconds) not to exceed.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Replace by the average of the previous and next values: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics of the selected series.

Charts tab:

Display charts: Activate this option to display the charts of the series before and after transformation.

Results

For each series, the following results are displayed:

Summary statistics: This table displays for the selected variables, the number of observations, the number of missing values, the number of non- missing values, the mean and the standard deviation (unbiased).

The results of various tests are then displayed. For the Pettitt's test, the SNHT the Buishand's Q test, charts are displayed with means μ_1 and μ_2 if a change-point is detected and μ if no change-point is detected.

Example

A tutorial explaining how to use the homogeneity tests is available at the XLSTAT Help Center. To consult the tutorial, please go to:

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

References

Alexandersson H. (1986). A homogeneity test applied to precipitation data. *Journal of Climatology* , **6**, 661-675.

Buishand T.A. (19 82). Some methods for testing the homogeneity of rainfall data. *Journal of Hydrology*, **58**, 11-27.

Pettitt A.N. (19 79). 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.

Durbin-Watson test

Use this tool to check if the residuals of a linear regression are autocorrelated.

In this section:

Description

Developed by J.Durbin and G.Watson (1950,1951), the Durbin-Watson test is used to detect the autocorrelation in the residuals from a linear regression.

Denote by Y the dependent variable, X the matrix of explanatory variables, α and β the coefficients abd ϵ the error term. Consider the following model:

$$y_t = \alpha + \beta x_t + \epsilon_t$$

In practice, the errors are often autocorrelated, it leads to undesirable consequences such as sub-optimal least-squares estimates. The Durbin-Watson test is used to detect autocorrelations in the error terms.

Assume that the $\{\epsilon_t\}$ are stationary and normally distributed with mean 0. The null and alternative hypotheses of the Durbin-Watson test are:

- H0: The errors are uncorrelated.
- Ha: The errors are $AR(p)$, where p is the order of autocorrelation.

$AR(p)$ is an autroregressive process of order p.

The Durbin-Watson D statistic writes:

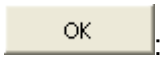
$$D = \frac{\sum_{t=p+1}^{n} (\epsilon_t - \epsilon_{t-p})^2}{\sum_{t=r+1}^{n} \epsilon_t^2}$$

In the context of the Durbin-Watson test, the main problem is the evaluation of the p-values which cannot be computed directly. XLSTAT uses the Pan (1968) algorithm for time series with

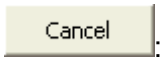
less than 70 observations and the Imhof(1961) procedure when there are more than 70 observations.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Residuals: Select the residuals from the linear regression. If the variable header has been selected, check that the "Variable labels" option has been activated.

X / Explanatory variables: Select the quantitative explanatory variables in the Excel worksheet. The data selected must be of numeric type. If the variable header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Options tab:

Significance level (%): Enter the significance level for the test (default value: 5%)

Order: Enter the order, i.e. the number of lags for the residuals (default value: 1)

Missing tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations which include missing data.

Replace by the average of the previous and next values: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

Ignore missing data: Activate this option to ignore missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the residuals.

Results

Summary statistics: The tables of descriptive statistics show the simple statistics for the residuals. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed.

The results of of the Durbin-Watson test are then displayed.

Example

A tutorial on the Durbin-Watson test is available on the XLSTAT Help Center:

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

References

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.

Cochrane-Orcutt estimation

Use this tool to account for serial correlation in the error term of a linear model.

In this section:

Description

Developed by D.Cochrane and G. Orcutt in 1949, the Cochrane-Orcutt estimation is a well-known econometric approach to account for serial correlation in the error term of a linear model. In case of serial correlation, usual linear regression is invalid because the standard errors are not unbiased.

Denote by Y the dependent variable, X the matrix of explanatory variables, α and β the coefficients and ϵ the error term. Consider the following model:

$$y_t = \alpha + \beta x_t + \epsilon_t$$

And suppose that the error term ϵ is generated by a stationary first-order autoregressive process such that:

$$\epsilon_t = \rho \epsilon_{t-1} + \epsilon_t, \text{ with } |\rho| < 1$$

with ϵ_t as a white noise.

To estimate the coefficients, the Cochrane-Orcutt procedure is based on the following transformed model:

$$\forall t \geq 2, y_t - \rho y_{t-1} = \alpha(1-\rho) + \beta(X_t - \rho X_{t-1}) + \epsilon_t$$

By introducing 3 new variables such as

$$Y^* = y_t - \rho y_{t-1}, X^* = X_t - \rho X_{t-1}, \lambda^* = 1 - \rho$$

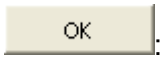
we have:

$$\forall t \geq 2, y_t^* = \alpha \lambda^* + \beta X_t^* + e_t$$

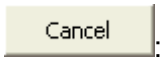
Since $\{e_t\}$ is a white noise, usual statistical inference can now be used.

Dialog box

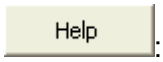
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Dependent variables:

Quantitative: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

X / Explanatory variables:

Quantitative: Select the quantitative explanatory variables in the Excel worksheet. The data selected must be of numeric type. If the variable header has been selected, check that the "Variable labels" option has been activated.

Date data: Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format. If this option is not activated, XLSTAT creates its own time variable ranging from 1 to the number of data.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will all be taken as 1. Weights must be greater than or equal to 0. A weight of 2 is equivalent to repeating the same observation twice. If a column header has been selected, check that the "Variable labels" option has been activated.

Regression weights: Activate this option if you want to carry out a weighted least squares regression. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Tolerance: Activate this option to prevent the OLS regression calculation algorithm taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

Validation tab:

Validation: Activate this option if you want to use a sub-sample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.

- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: Activate this option if you want to select data to use them in prediction mode. If you activate this option, you need to make sure that the prediction dataset is structured as the estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

X / Explanatory variables: Select the quantitative explanatory variables. The first row must not include variable labels.

Observations labels: activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

- **Check for each Y separately**: Choose this option to remove the observations with missing data in the select Y (dependent) variables, only if the Y of interest has a missing data.
- **Across all Ys**: Choose this option to remove the observations with missing data in the Y (dependent) variables, even if the Y of interest has no missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the correlation matrix for quantitative variables (dependent or explanatory).

Analysis of variance: Activate this option to display the analysis of variance table.

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

Charts tab:

Regression charts: Activate this option to display regression chart.

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.
- **Predictions and residuals**: Activate this option to display the following charts.

(1) Line of regression: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

(2) Explanatory variable versus standardized residuals: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

(3) Dependent variable versus standardized residuals.

(4) Predictions for the dependent variable versus the dependent variable.

(5) Bar chart of standardized residuals.

Confidence intervals: Activate this option to have confidence intervals displayed on charts (1) and (4).

Results

Summary statistics: The tables of descriptive statistics show the simple statistics for all the variables selected. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed for the dependent variables (in blue) and the quantitative explanatory variables. For qualitative explanatory variables the names of the various categories are displayed together with their respective frequencies.

Correlation matrix: This table is displayed to give you a view of the correlations between the various variables selected.

Summary of the variables selection: Where a selection method has been chosen, XLSTAT displays the selection summary. For a stepwise selection, the statistics corresponding to the different steps are displayed. Where the best model for a number of variables varying from p to q has been selected, the best model for each number or variables is displayed with the corresponding statistics and the best model for the criterion chosen is displayed in bold.

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.
- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).
- **R²**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ with } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

The R^2 is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer R^2 is to 1, the better is the model. The problem with the R^2 is that it does not take into account the number of variables used to fit the model.

- **Adjusted R²**: The adjusted determination coefficient for the model. The adjusted R^2 can be negative if the R^2 is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}$$

The adjusted R^2 is a correction to the R^2 which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W - p^*}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.
- **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows:

$$MAPE = \frac{100}{W}\sum_{i=1}^{n} w_i \left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

- **DW**: The Durbin-Watson statistic is defined by:

$$DW = \frac{\sum_{i=2}^{n}\left[(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})\right]^2}{\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2}$$

This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the

basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- **Cp**: Mallows Cp coefficient is defined by:

$$Cp = \frac{SSE}{\hat{\sigma}} + 2p^* - W$$

where SSE is the sum of the squares of the errors for the model with p explanatory variables and $\hat{\sigma}$ is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to p*, the less the model is biased.

- **AIC**: Akaike's Information Criterion is defined by:

$$AIC = W \ln(\frac{SSE}{W}) + 2p^*$$

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz's Bayesian Criterion is defined by:

$$SBC = W \ln(\frac{SSE}{W}) + \ln(W)p^*$$

This criterion, proposed by Schwarz (1978) is similar to the AIC, and the aim is to minimize it.

- **PC**: Amemiya's Prediction Criterion is defined by:

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}$$

This criterion, proposed by Amemiya (1980) is used, like the adjusted R² to take account of the parsimony of the model.

- **Press RMSE**: Press' statistic is only displayed if the corresponding option has been activated in the dialog box. It is defined by:

$$Press = \sum_{i=1}^{n} w_i \left(y_i - \hat{y}_{i(-i)}\right)^2$$

where $\hat{y}_{i(-i)}$is the prediction for observation i when the latter is not used for estimating parameters. We then get:

$$\text{Press } RMSE = \sqrt{\frac{Press}{W - p^*}}$$

Press's RMSE can then be compared to the RMSE. A large difference between the two shows that the model is sensitive to the presence or absence of certain observations in the model.

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

The **parameters of the model** table display the estimate of the parameters, the corresponding standard error, the Student's t, the corresponding probability, as well as the confidence interval. The autocorrelation coefficient r is also displayed.

The **equation of the model** is then displayed to make it easier to read or re-use the model.

Autocorrelation coefficient: The estimated value of the autocorrelation coefficient r.

The table of **standardized coefficients** (also called beta coefficients) is used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of normalized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the residuals and the confidence intervals with the fitted prediction. Two types of confidence intervals are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the random values being larger. If the validation data have been selected, they are displayed at the end of the table.

The **charts** which follow show the results mentioned above. If there is only one explanatory variable in the model, the first chart displayed shows the observed values, the regression line and both types of confidence interval around the predictions. The second chart shows the normalized residuals as a function of the explanatory variable. In principle, the residuals should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a problem with the model.

The **three charts** displayed next show respectively the evolution of the standardized residuals as a function of the dependent variable, the distance between the predictions and the observations (for an ideal model, the points would all be on the bisector), and the standardized residuals on a bar chart. The last chart quickly shows if an abnormal number of values are outside the interval]-2, 2[given that the latter, assuming that the sample is normally distributed, should contain about 95% of the data.

If you have selected the data to be used for calculating **predictions on new observations**, the corresponding table is displayed next.

Example

A tutorial on the Cochrane-Orcutt estimation is available on XLSTAT Help Center:

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

References

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

Heteroscedasticity tests

Use this tool to determine whether the residuals from a linear regression can be considered as having a variance that is independent of the observations.

In this section:

Description

The concept of heteroscedasticity - the opposite being homoscedasticity - is used in statistics, especially in the context of linear regression or for time series analysis, to describe the case where the variance of errors or the model is not the same for all observations, while often one of the basic assumption in modeling is that the variances are homogeneous and that the errors of the model are identically distributed.

In linear regression analysis, the fact that the errors of the model (also named residuals) are not homoskedastic has the consequence that the model coefficients estimated using ordinary least squares (OLS) are neither unbiased nor those with minimum variance. The estimation of their variance is not reliable.

If it is suspected that the variances are not homogeneous (a representation of the residuals against the explanatory variables may reveal heteroscedasticity), it is necessary to perform a test for heteroscedasticity. Several tests have been developed, with the following null and alternative hypotheses:

- H0 : The residuals are homoscedastic
- Ha : The residuals are heteroscedastic

Breusch-Pagan test

This test has been developed by Breusch and Pagan (1979), and later improved by Koenker (1981) - which is why this test is sometimes named the Breusch- Pagan and Koenker test - to allow identifying cases of heteroscedasticity, which make the classical estimators of the parameters of the linear regression unreliable. If e is the vector of the errors of the model, the null hypothesis H0 can write:

$$H0 : Var(u/x) = \sigma^2$$

$$H0 : Var(u/x) = E(e^2/x) = E(e^2/x_1, x_2, ..., x_k) = E(e^2) = \sigma^2$$

To verify that the quadratic errors are independent of the explanatory variables, which can translate into many functional forms, the simplest is to regress the squared errors by the explanatory variables. If the data are homoskedastic, the coefficient of determination R^2 should then not be equal to 0. If H0 is not rejected we can conclude that heteroscedasticity, if it exists, does not take the functional form used. Practice shows that heteroscedasticity is not a problem if H0 is not accepted. If H0 is rejected, it is likely that there is heteroscedasticity and that it takes the functional form described above.

The statistic used for the test, proposed by Koenker (1981) is:

$$LM = nR^2$$

where LM stands for *Lagrange multiplier*. This statistic has the advantage of asymptotically following a Chi-square distribution with p degrees of freedom, where p is the number of explanatory variables.

If the null hypothesis is rejected, it will be necessary to transform the data before doing the regression, or using modeling methods to take into account the variability of the variance.

White test and modified Whitetest (Wooldridge)

This test was developed by White (1980) to identify cases of heteroscedasticity making classical estimators of the parameters of linear regression unreliable. The idea is similar to that of Breusch and Pagan, but it relies on weaker assumptions as for the form that heteroscedasticity takes. This results in a regression of the quadratic errors by the explanatory variables and by the squares and cross-products of the latter (for example for two regressors, we take x1, x2, x1², x2², x1x2 to model squared errors). The statistic used is the same as the test-Breusch Pagan, but due to the presence of many more regressors, there are here $2p + p * (p - 1)/2$ degrees of freedom for the Chi-square distribution.

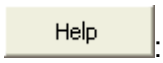
In order to avoid losing too many degrees of freedom, Wooldrigde (2009) proposed to regress the squared errors by the model predictions and by their square. This reduces to 2 the number of degrees of freedom for the Chi-square.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Residuals: Select the residuals from the linear regression. If the variable header has been selected, check that the "Labels included" option has been activated.

X / Explanatory variables: Select the quantitative explanatory variables in the Excel worksheet. The data selected must be of numeric type. If the variable header has been selected, check that the "Labels included" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Breusch-Pagan test: Activate this option to run a Breusch-Pagan test.

White test: Activate this option to run a White test. Activate the "Wooldridge" option if you want to use the modified version of the test (see the description chapter for further details).

Options tab:

Significance level (%): Enter the significance level for the test (default value: 5%).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics of the selected series.

Charts tab:

Display charts: Activate this option to display the scatter plot of the residuals versus the explanatory variable.

Results

Summary statistics: This table displays for the selected variables, the number of observations, the number of missing values, the number of non- missing values, the mean and the standard deviation (unbiased).

The results of of the selected tests are then displayed.

Example

A tutorial explaining how to use the heteroscedasticity tests is available on the XLSTAT Help Center. To consult the tutorial, please go to:

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

References

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.

Unit root and stationarity tests

Use this tool to determine whether a series is stationary or not.

In this section:

Description

A time series Y_t $(t = 1, 2...)$ is said to be stationary (in the week sense) if its statistical properties do not vary with time (expectation, variance, autocorrelation). The white noise is an example of a stationary time series, with for example the case where Y_t follows a normal distribution $N(\mu, \sigma^2)$ independent of t. An example of a non-stationary series is the random walk defined by:

$$Y_t = Y_{t-1} + \epsilon_t$$

where ϵ_t is a white noise.

Identifying that a series is not stationary allows to afterwards study where the non-stationarity comes from. A non-stationary series can, for example, be **stationary in difference**: Y_t is not stationary, but the $Y_t - Y_{t-1}$ difference is stationary. It is the case of the random walk. A series can also be **stationary in trend**. It is the case with the series defined by:

$$Y_t = 0.5X_{t-1} + 1.4t + \epsilon_t$$

where ϵ_t is a white noise, that is not stationary.

On the other hand, the series

$$Y_t - 1.4t = 0.5Y_{t-1} + \epsilon_t$$

is stationary. Y_t is also stationary in difference.

Stationarity tests allow verifying whether a series is stationary or not. There are two different approaches: some tests consider as null hypothesis H0 that the series is stationary (KPSS test, Leybourne and McCabe test), and for other tests, on the opposite, the null hypothesis is on the contrary that the series is not stationary (Dickey-Fuller test, augmented Dickey-Fuller test,

Phillips-Perron test, DF-GLS test). XLSTAT includes as of today the KPSS and Dickey-Fuller tests.

Dickey-Fuller test

This test has been developed by Dickey and Fuller (1979) to allow identifying a unit root in a time series for which one thinks there is an order 1 autoregressive component, and may be as well a trend component linearly related to the time. As a reminder, an order 1 autoregressive model (noted AR(1)), can be written as follows:

$$X_t = \rho X_{t-1} + \epsilon_t \ , t = 1, 2, \ldots$$

where the ϵ_t are independent identically distributed variables that follow an $N(0, \sigma^2)$ normal distribution.

The series is stationary if $|\rho| < 1$. It is not stationary and corresponds to a random walk if $\rho = 1$.

If one adds a constant and a trend to the model, the model writes:

$$X_t = \rho X_{t-1} + \alpha + \beta t + \epsilon_t \ , t = 1, 2....$$

where the ϵ_t are independent identically distributed variables that follow an $N(0, \sigma^2)$ normal distribution.

Dickey and Fuller decided to take as null hypothesis ρ =1 because it has an immediate operational impact: if the null hypothesis is not rejected, then, in order to be able to analyze the time series and if necessary to make predictions, it is necessary to transform the series, using differencing (see the Time series transformation and ARIMA tools).

The two possible alternative hypotheses are:

Ha(1): $|\rho| < 1$, the series is stationary

Ha(2): $|\rho| > 1$, the series is explosive

The statistics used in the Dickey-Fuller test are computed using a linear regression model, and correspond to the *t* statistic computed by dividing the coefficient of the model by its standard error. Dickey and Fuller define:

- AR(1) model:

$$\hat{\tau} = (\hat{\rho} - 1) / \sqrt{S_1^2 c_1}$$

- AR(1) model with constant μ:

$$\hat{\tau}_\mu = (\hat{\rho}_\mu - 1) / \sqrt{S_2^2 c_2}$$

- AR(1) model with constant μ and a linear trend function of t:

$$\hat{\tau}_\tau = (\hat{\rho}_\tau - 1)/\sqrt{S_3^2 c_3}$$

The S_k^2 correspond to the mean squared error and the c_k to variances.

While these statistics are straightforward to compute, their exact and asymptotic distributions are complex. The critical values have been estimated through Monte Carlo simulations by the authors, with several improvements over time, as the machines allowed more simulations. MacKinnon (1996) has proposed an approach based on numerous Monte Carlo simulations that allows to compute p-values and critical values for various sample sizes.XLSTAT estimates critical values and p-values either by running a predefined set of Monte Carlo simulations for the considered sample size or the surface regression approach proposed by MacKinnon (1996).

Dickey et Fuller have shown that these distributions do not depend on the distribution of the ϵ_t and on the initial value of the series, Y_0.

Fuller (1976) had already shown that this approach can be generalized to AR(p) models to determine whether there exists a unit root while not being able to say from which term in the model the non-stationarity comes from.

Augmented Dickey-Fuller test

This test has been developed by Said et Dickey (1984) and complements the Dickey-Fuller test by generalizing the approach valid for AR(p) models to ARMA(p, q) models, for which we assume that it is in fact an ARIMA(p, d, q), with d=1 under the null hypothesis H0. Said and Dickey show that it is not necessary to know p, d and q to apply the Dickey-Fuller test presented above. However, a k parameter, corresponding to the horizon to consider for the mobile mean of the model must be provided by the user so that the test can be run. By default, XLSTAT recommends the following value:

$$k = INT((n-1)^{1/3})$$

where $INT()$ is the integer part

Said and Dickey show that the statistic t of the Dickey-Fuller test can be used. Its asymptotic distribution is the same as the one of the Dickey-Fuller test.

Phillips-Perron test

An alternative generalization of the Dickey-Fuller test to more complex data generation processes was introduced by Phillips (1987a) and further developed in Perron (1988) and Phillips and Perron (1988).

As for the DF test, three possible regressions are considered in the Phillips- Perron (PP) test, namely, without an intercept, with an intercept and with an intercept and a time trend. Those are given in the following equations, respectively.

$$X_t = \rho X_{t-1} + \epsilon_t$$
$$X_t = \rho X_{t-1} + \alpha + \epsilon_t$$
$$X_t = \rho X_{t-1} + \alpha + \beta \cdot (t - T/2) + \epsilon_t$$

It should be noted that within the PP test, the error term ϵ_t is expected to have a null average but it can be serially correlated and/or heteroscedastic.

Unlike the augmented Dickey-Fuller (ADF) test, the PP test does not deal with serial correlation at the regression level. Instead, a non parametric correction is applied to the statistic itself to account for potential effects of heteroscedasticity and serial correlations on the adjustment residuals. The statistic noted Z_t is given by:

$$Z_t = \frac{\hat{\sigma}}{\hat{\lambda}} t_\rho - \frac{1}{2} \left(\frac{\hat{\lambda^2} - \hat{\sigma^2}}{\hat{\lambda}^2} \right) \left(\frac{T \times SE(\hat{\rho})}{\hat{\sigma}^2} \right)$$

Where $\hat{\lambda}^2$ and $\hat{\sigma}^2$ are consistent estimators of the variance parameters:

$$\hat{\lambda}^2 = \lim_{x \to +\infty} T^{-1} \sum_t^T E \left[T^{-1} (\sum_{t=1}^r \epsilon_t)^2 \right]$$

$$\hat{\sigma^2} = \lim_{x \to +\infty} T^{-1} \sum_t^T E[\epsilon_t^2]$$

and

$$t_\rho = \frac{\hat{\rho} - 1}{SE(\hat{\rho})}$$

The estimator $\hat{\lambda^2}$ is the one proposed by Newey and West (1987). It guarantees the robustness of the statisic against heteroscedasticity and serial correlations.

- Short (default option): the number of steps considered for the computation of the Newey-West estimator is given by:

$$k = INT(4.(\frac{T}{100})^{2/9})$$

- Long : for series resulting from a higher-order MA process, the number of steps is given by

$$k = INT(12(\frac{T}{100})^{2/9})$$

Where $INT()$ is the integer part.

The PP test uses the same distribution as the DF or ADF t-statistic. Critical value and p-value estimates are made following the surface regression approach proposed by MacKinnon (1996)

or using Monte Carlo simulations.

One of the advantages of the PP test over the ADF test is to allow for heteroscedasticity in the data generation process of ϵ_t. Furthermore, there is no need for a sensitive parametrization of the Newey-West estimator as for the ADF test.

KPSS test of stationarity

This test takes its name from its authors, Kwiatkowski, Phillips, Schmidt and Shin (1991). Contrary to the Dickey-Fuller tests, this test allows testing the null hypothesis that the series is stationary. Consider the model where

$$Y_t = \xi t + r_t + \epsilon_t \ , t = 1, 2, ...$$

where ϵ_t is a stationary error, and r_t is a random walk defined by

$$r_t = r_{t-1} + u_t,$$

where r_0 is a constant, and the u_t are independent identically distributed variables with mean 0 and variance σ^2.

The Y_t series is stationary in the case where the σ^2 variance is null. It is stationary in trend if ξ is not null, and stationary in level (around r_0) if $\xi = 0$.

Let n be the number of time steps available for the series. Let e_t be the residuals, when regressing the y_t by the time and a constant, when one wants to test stationarity in trend, or when comparing the series with its mean when testing for stationarity in level.

We define:

$$s^2(l) = \frac{1}{n}\sum_{t=1}^{n} e_t^2 + \frac{2}{n}\sum_{s=1}^{l} w(s,l) \sum_{t=s+1}^{n} e_t e_{t-s}$$

with

$$w(s,l) = 1 - s(l+1)$$

Let S_t^2 be the mean of squared errors between times 1 and t. The statistic used for the "Level" stationarity test is given by:

$$\eta_\mu = \frac{1}{n^2}\sum_{t=1}^{n} S_t^2 / s^2(l)$$

For the "Trend" stationarity test we use:

$$\eta_\tau = \frac{1}{n^2}\sum_{t=1}^{n} S_t^2 / s^2(l)$$

the difference between both comes from the different residuals.

As with the Dickey-Fuller test, these statistics are easy to compute, but their exact and asymptotic distributions are complex. Kwiatkowski *et al.* computed the asymptotic critical values using Monte Carlo simulations. XLSTAT allows to compute critical values and p-values adapted to the size of the sample, using Monte Carlo simulations for each new run.

Weighting with the Newey-West method

The Newey-West (1987) estimator is used to reduce the effect of dependence (correlation, autocorrelation) and heteroscedasticity (non homogeneous variances) of the error terms of a model. The idea is to balance the model errors in the calculation of statistics involving them. If L is the number of steps taken into account, the weight of each error is given by:

$$w_l = 1 - \frac{l}{L+1}\ , l = 1, 2..., L$$

The KPSS test uses linear regressions that assume the homoscedasticity of the errors. The use of the Newey-West weighting is recommended by the authors and is available in XLSTAT. XLSTAT recommends for the value of L:

- Short: $L = INT(3 * \sqrt{n}/13)$

- Long: $L = INT(10^* \sqrt{n}/14)$

where $INT()$ is the integer part.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

General tab:

Times series: Select the data that correspond to the time series. If a header is available on the first row, make sure you activate the "Series labels" option.

Date data: Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format. If this option is not activated, XLSTAT creates its own time variable ranging from 1 to the number of data.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Series labels: Activate this option if the first row of the selected series includes a header.

Dicker-Fuller test: Activate this option to run a Dickey-Fuller test. Choose the type of test you want to use (see the description section for further details).

Phillips-Perron test: Activate this option to run a Phillips-Perron test. Choose the type of test you want to use (see the description section for further details).

KPSS test: Activate this option to run a KPSS test. Choose the type of test you want to use (see the description section for further details).

Options tab:

Significance level (%): Enter the significance level for the test (default value: 5%).

Method: choose the method to use for the p-value and critical value estimates

- Surface regression: selects the approach proposed by MacKinnon (1996).
- Monte Carlo: selects Monte Carlo simulations based estimates.

Dickey-Fuller test: In the case of a Dickey-Fuller test, you can use the default value of k (see the "Description" section for more details) or enter your own value.

Phillips-Perron test: for a Phillips-Perron test, you should select either the short (default value) or the long number of steps (see the "Description" section for more details).

KPSS test: Choose whether you want to use the Newey-West to remove the impact of possible autocorrelations in the residuals of the model. For the lag to apply, you can choose between short, long, or you can enter your own value for L (see the "Description" section for more details).

Missing tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Replace by the average of the previous and next values: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

Ignore missing data: Activate this option to ignore missing data.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics of the selected series.

Charts tab:

Display charts: Activate this option to display the charts of the series.

Results

Summary statistics: This table displays for the selected variables, the number of observations, the number of missing values, the number of non- missing values, the mean and the standard deviation (unbiased).

The results of of the selected tests are then displayed.

Example

A tutorial explaining how to perform unit root tests or stationarity test is available on the XLSTAT Help Center. To consult the tutorial, please go to:

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

References

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.

Cointegration tests

Use this module to perform VAR-based cointegration tests on a group of two or more I(1) time series using the approach proposed by Johansen (1991, 1995).

In this section:

Description

Economic theory often suggests long term relationship between two or more economic variables. Although those variables can derive from each other on a short term basis, the economic forces at work should restore the original equilibrium between them in the long run. Examples of such relationships in economics include money with income, prices and interest rates or exchange rate with foreign and domestic prices. In finance, such relationships are expected for instance between the prices of the same asset on different market places.

The term of cointegration was first introduced by Engle and Granger (1987) after the work of Granger and Newbold (1974) on spurious regression. It identifies a situation where two or more non stationary time series are bound together in such a way that they cannot deviate from some equilibrium in the long term. In other words, there exists one or more linear combination of those I(1) time series (or integrated of order 1, see unit root test) that is stationary (or I(0)). Those stationary combinations are called cointegrating equations.

One of the most interesting approaches for testing for cointegration within a group of time series is the maximum likelihood methodology proposed by Johansen (1988, 1991). This approach, implemented in XLSTAT, is based on Vector Autoregressive (VAR) models and can be described as follows.

First consider the levels $VAR(P)$ model for Y_t, a K-vector of I(1) time series:

$$Y_t = \Phi D_t + \Pi_1 Y_{t-1} + ... + \Pi_P Y_{t-P} + \epsilon_t \text{ for } t = 1, ..., T$$

Where D_t contains deterministic terms such as constant or trend and ϵ_t is the vector of innovations.

The parameter P is the VAR order and is one of the input parameter to Johansen's methodology for testing cointegration. If you don't know which value this parameter should take for you data set, you should select the option automatic in the General tab. You will then have to specify the model that best describes your data in the option tab (no trend nor intercept,

intercept, trend or trend and intercept), set a maximum number of lags to evaluate and choose the discriminating criterion among the 4 proposed (AIC, FPE, HQ, BIC). XLSTAT will then estimate the parameter P following the approach detailed in Lüktepohl (2005) and perform subsequent analysis. Detailed results are provided at the end of the analysis for further control.

According to the Granger representation theorem, a $VAR(P)$ model with I(1) variables can equivalently be represented as a Vector Error Correction Model (VECM):

$$\Delta Y_t = \Phi D_t + \Pi Y_{t-1} + \Gamma_1 \Delta Y_{t-1} + ... + \Gamma_{P-1} \Delta Y_{t-P+1} + \epsilon_t$$

Where Δ denotes the difference operator, $\Pi = \Pi_1 + ... + \Pi_{P-1} - I_K$ and $\Gamma_l = -\sum_{j=l+1}^{P} \Pi_j$ for $l = 1, ..., P-1$.

In this representation, ΔY_t and its lags are all I(0). The term Y_{t-1} is the only potentially non stationary component. Therefore for the above equation to hold (a linear combination of I(0) terms is also I(0)), the term $\Pi Y_{t=-1}$ must contain the cointegration relationship if it exists.

Three cases can be considered:

- the matrix Π equals 0 ($rank(\Pi) = 0$), then no cointegration exists,
- the matrix Π has full rank ($rank(\Pi) = K$), then each independent component of Y_{t-1} is I(0) (which violates or first assumption of I(1) series),
- the matrix Π is neither null nor of full rank ($0 < rank(\Pi) < K$), then Y_{t-1} is I(1) with r linearly independent cointegrating vectors and $K - r$ common stochastic trends.

In the latter case, the matrix Π can be written as the product:

$$\Pi_{(K \times K)} = \alpha_{(K \times r)} \beta_{(r \times K)}$$

Where $rank(\alpha) = rank(\beta) = r$.

The matrix β is the cointegrating matrix and its columns form a basis for the cointegrating coefficients. The matrix α also known as the adjustment matrix (or loading matrix) controls the speed at witch the effect of Y_{t-1} propagates toΔY_t. It is important to note that the factorization $\Pi = \alpha\beta'$ is not uniquely defined and may require some arbitrary normalization to obtain unique values of α andβ. Values reported in XLSTAT use the normalization $\beta' \cdot S_{11} \cdot \beta = I_r$ proposed by Johansen (1995).

The test methodology estimates the matrix Π and constructs successive likelihood ratio (LR) tests for its reduced rank on the estimated eigenvalues ofΠ: $\widehat{\lambda_1} > \widehat{\lambda_2} > \ldots > \widehat{\lambda_x}$. The reduced rank of Π is equal to the number of non-zero eigenvalues. It is also the rank of cointegration of the system (or equivalently the number of cointegrating equations).

Two sequential procedures proposed by Johansen are implemented to evaluate the cointegration rank r_0:

- the λ_{max}-test (or lambda max) uses the statistic $LR_{max}(r_0) = -T \ln(1 - \hat{\lambda}_{r_0+1})$,

- the trace test for which the statistic is $LR_{trace}(r_0) = -T\sum_{i=r_0+1}^{n}\ln(1-\hat{\lambda}_i)$.

Starting from the Null hypothesis that non cointegration relationship exists ($r_0 = 0$), the λ_{max}-test will test that the $(r_0+1)^{th}$ eigenvalue can be accepted to be zero. If the hypothesis of $\lambda_{r_0+1} \approx 0$ is rejected, then the next level of cointegration can be tested. Similarly, LR_{trace} of the trace test should be close to zero if the rank of Π equals r_0 and large if it is greater than r_0.

The asymptotic distributions of those LR tests are non standard and depend on the assumption made on the deterministic trends of ΔY_t which can be rewritten as:

$$\Delta Y_t = c_1 + d_1 t + \alpha(\beta' Y_{t-1} + c_0 + d_0 t) + \Gamma_1 \Delta Y_{t-1} + ... + \Gamma_{P-1}\Delta Y_{t-P+1} + \epsilon_t$$

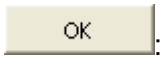
5 types of restriction are considered depending on the trending nature of both Y_t and $\beta' Y_t$ (the cointegrating relationships):

- **H2** ($c_0 = c_1 = d_0 = d_1 = 0$): the series in Y_t are I(1) with no deterministic trends in levels and $\beta' Y_t$ have means zero. In practice, this case is rarely used.
- **H1*** ($c_1 = d_0 = d_1 = 0$): the series in Y_t are I(1) with no deterministic trends in levels and $\beta' Y_t$ have non-zero means.
- **H1** ($d_0 = d_1 = 0$): the series in Y_t are I(1) with linear trends in levels and $\beta' Y_t$ have non-zero means.
- **H*** ($d_1 = 0$): the series in Y_t and $\beta' Y_t$ have linear trends.
- **H** (unconstrained): the series in Y_t are I(1) with quadratic trends in levels and $\beta' Y_t$ have linear trends. Again, this case is hardly used in practice.

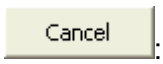
To perform a cointegration test in XLSTAT, you have to choose one of the above assumptions. The choice should be motivated by the specific nature of your data and the considered economics model. However, if it is unclear which restriction applies best, a good strategy might be to evaluate the robustness of the result by successively selecting a different assumption among **H1***, **H1** and **H*** (the remaining 2 options being very specific and easily identifiable). Critical values and p-values for both the λ_{max}-test and the trace test are computed in XLSTAT as proposed by MacKinnon-Haug-Mechelis (1998).

Dialog box

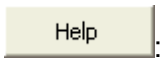
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

General tab:

Times series: Select the data that correspond to the time series. If a header is available on the first row, make sure you activate the "Series labels" option.

Date data: Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format. If this option is not activated, XLSTAT creates its own time variable ranging from 1 to the number of data.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Series labels: Activate this option if the first row of the selected series includes a header.

Model: Select between **H2**, **H1***, **H1**, **H*** and H the type of restriction that best describes your data set (see the description for further details).

VAR order: Select the automatic option for an automatic estimation of the P parameter (see the description for further details) or select the user defined option and enter your own value.

Options tab:

Significance level (%): Enter the significance level for the test (default value: 5%).

VAR order estimation: If the automatic option is selected for the VAR order on the General tab, you must set three parameters: the model, the selection criterion and the maximum number of lag.

Model: Select between None, Intercept, Trend and Intercept + trend the model that best describes your time series.

Selection criterion: Select between the four criteria computed (AIC, FPE, HQ and BIC), the one XLSTAT will use to select the VAR order.

Maximum number of lag: Select the maximum number of lag that will be computed by XLSTAT to select the VAR order.

Missing tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Replace by the average of the previous and next values: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics of the selected series.

Results

Summary statistics: This table displays for the selected variables, the number of observations, the number of missing values, the number of non- missing values, the mean and the standard deviation (unbiased).

VAR order estimation: If the automatic option is selected for the VAR order, this table displays the four criteria values for the VAR order estimation. Each line corresponds to the evaluation of one number of lags from 1 up to the maximum number of lag. The discriminating criterion is in bold.

Lambda max test: This table displays for each rank of cointegration tested the corresponding eigenvalue, the lambda max test statistic and the associated critical value and p-values.

Trace test: This table displays for each rank of cointegration tested the corresponding eigenvalue, the trace test statistic and the associated critical value and p-values.

Adjustment coefficients (alpha): This table displays the resulting loading matrix α(see description for further details).

Cointegration coefficients (beta): This table displays the cointegrating matrix β(see description for further details).

Example

A tutorial explaining how to perform cointegration analysis on time series is available on the XLSTAT Help Center. To consult the tutorial, please go to:

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

References

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

Time series transformation

Use this tool to transform a time series A into a time series B that has better properties: removed trend, reduced seasonality, and better normality.

In this section:

Description

XLSTAT offers four different possibilities for transforming a time series X_t into Y_t, $(t = 1, \ldots, n)$:

Box-Cox transformation, to improve the normality of the time series; the Box-Cox transformation is defined by the following equation:

$$Y_t = \begin{cases} \frac{X_t^{\lambda}-1}{\lambda}, & (X_t > 0,\ \lambda \neq 0)\ or\ (X_t \geq 0,\ \lambda > 0) \\ ln(X_t), & (X_t > 0, \lambda = 0) \end{cases}$$

XLSTAT accepts a fixed value of λ, or it can find the value that maximizes the likelihood of the residuals, the model being a simple linear model with the time as sole explanatory variable.

Differencing, to remove trend and seasonalities and to obtain stationarity of the time series. The difference equation writes:

$$Y_t = (1 - B)^d (1 - B^s)^D X_t$$

where d is the order of the first differencing component, s is the period of the seasonal component, D is the order of the seasonal component, and B is the lag operator defined by:

$$BX_t = X_{t-1}$$

The values of (d, D, s) can be chosen in a trial and error process, or guessed by looking at the descriptive functions (ACF, PACF). Typical values are $(1, 1, s)$, $(2, 1, s)$. s is 12 for monthly data with a yearly seasonality, 0 when there is no seasonality.

Detrending and **deseasonalizing**, using the classical decomposition model which writes:

$$X_t = m_t + s_t + \epsilon_t$$

wherem_t is the trend component and s_t the seasonal component, and ϵ_t is a $N(0,1)$ white noise component. XLSTAT allows to fit this model in two separate and/or successive steps:

1 – Detrending model:

$$X_t = m_t + \epsilon_t = \sum_{i=0}^{k} a_i t^i + \epsilon_t$$

where k is the polynomial degree. The a_i parameters are obtained by fitting a linear model to the data. The transformed time series writes:

$$Y_t = \epsilon_t = X_t - \sum_{i=0}^{p} a_i t^i$$

2 – Deseasonalization model:

$$X_t = s_t + \epsilon_t = \mu + b_i + \epsilon_t, \quad i = t \bmod p$$

where p is the period. The b_i parameters are obtained by fitting a linear model to the data. The transformed time series writes:

$$Y_t = \epsilon_t = X_t - b_i - \mu$$

Note: there exist many other possible transformations. Some of them are available in the Variables transformations feature (see the "Preparing data" section). Linear filters may also be applied. Moving average smoothing methods which are linear filters are available in the "Smoothing" tool of XLSTAT.

Seasonal decomposition: from a user-defined period P, the seasonal decomposition estimates and decomposes the time series into 3 components (trend, seasonal and random).

If the chosen model type is additive, the model can be expressed as follows:

$$X_t = m_t + s_{t \bmod p} + \epsilon_t$$

with X_t the initial time series, m_t the trend component, $s_{t \bmod p}$ the seasonal component and ϵ_t the random component.

First, the trend component is estimated by applying a centered moving average filter to X_t:

$$\hat{m_t} = \sum_{i=-P/2}^{P/2} W_i \times X_{t+i}$$

where $P/2$ is the integer division of P by 2 and the coefficients w_i are defined as follows:

$$W_i = \begin{cases} \frac{1}{2P}, & if \mid i \mid = P/2 \\ \\ \frac{1}{P}, & otherwise \end{cases}$$

Each seasonal index s_i is computed from the difference $s_t = X_t - m_t$ as the average of the elements of s_t for which $t \bmod P = i$

Their values are then centered as shown below:

$$\hat{S}_i = \hat{s}_i - \frac{1}{p}\sum_{j=1}^{p} \hat{s}_j$$

Finally, the random component is estimated as follows:

$$\hat{\epsilon}_i = X_t - \hat{m}_t - \hat{s}_{t \bmod P}$$

If the multiplicative type of decomposition is chosen, the model is given by:

$$X_t = m_t \times s_{t \bmod P} \times \epsilon_t$$

The trend component is estimated in the same way as indicated for the additive decomposition.

The seasonal indices s_i are computed as the average of the elements of $s_t = X_t / m_t$ for which $t \bmod P = i$.

They are then normalized as follows:

$$\hat{S}_i = \hat{s}_i \times (\prod_{j=1}^{p} \hat{s}_j)^{-1/P}$$

Finally, the estimated random component is given by:

$$\hat{\epsilon_i} = \frac{X_t}{\hat{s}_{t \bmod P} \times \hat{m}_t}$$

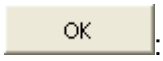
The Empirical Mode Decomposition (EMD) enables one to decompose any complex signal into a finite and often small number of components called IMFs, for intrinsic mode functions. By performing a Hilbert Transform on them, these components yield instantaneous frequencies as functions of time, which gives them a relevant physical interpretation.

The advantages of this method are the following: * It can be used on non-stationnary and nonlinear time series * It is adaptive, as the components are deduced from the data itself, and therefore quite efficient

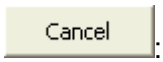
The **Ensemble Empirical Mode Decomposition** (EEMD) was invented in 2009 as a Noise Assisted Data Analysis (NADA) method, and intended to answer the mode mixing issue which was sometimes encountered when using the original EMD algorithm. Indeed, when performing a regular EMD on a signal that contains intermittent components, an IMF would sometimes represent several modes, or time scales, in it.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Times series: Select the data that correspond to the time series. If a header is available on the first row, make sure you activate the "Series labels" option.

Date data: Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format. If this option is not activated, XLSTAT creates its own time variable ranging from 1 to the number of data.

- **Check intervals**: Activate this option so that XLSTAT checks that the spacing between the date data is regular.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Series labels: Activate this option if the first row of the selected series includes a header.

Options tab:

Box-Cox: Activate this option to compute the Box-Cox transformation. You can either fix the value of the Lambda parameter, or decide to let XLSTAT optimize it (see the description for further details).

Differencing: Activate this option to compute differenced series. You need to enter the differencing orders (d, D, s). See the description for further details.

Polynomial regression: Activate this option to detrend the time series. You need to enter polynomial degree. See the description for further details.

Deseasonalization: Activate this option to remove the seasonal components using a linear model. You need to enter the period of the series. See the description for further details.

Seasonal decomposition: Activate this option to compute the seasonal indices and decompose the time series. You need to select a model type, additive or multiplicative and enter the period of the series. See the description for further details.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

- **Check for each Y separately**: Choose this option to remove the observations with missing data in the select Y (dependent) variables, only if the Y of interest has a missing data.
- **Across all Ys**: Choose this option to remove the observations with missing data in the Y (dependent) variables, even if the Y of interest has no missing data.

Replace by the average of the previous and next values: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics of the selected series.

Charts tab:

Display charts: Activate this option to display the charts of the series before and after transformation.

Results

Summary statistics: This table displays for the selected variables, the number of observations, the number of missing values, the number of non- missing values, the mean and the standard deviation (unbiased).

Box-Cox transformation:

Estimates of the parameters of the model: This table is available only if the Lambda parameter has been optimized. It displays the three parameters of the model, which are Lambda, the Intercept of the model and slope coefficient.

Series before and after transformation: This table displays the series before and after transformation. If Lambda has been optimized, the transformed series corresponds to the residuals of the model. If it hasn't then the transformed series is the direct application of the Box-Cox transformation.

Differencing

Series before and after transformation: This table displays the series before transformation and the differenced series. The first $d + D + s$ data are not available in the transformed series because of the lag due to the differencing itself.

Detrending (Polynomial regression)

Goodness of fit coefficients: This table displays the goodness of fit coefficients.

Estimates of the parameters of the model: This table displays the parameters of the model.

Series before and after transformation: This table displays the series before and after transformation. The transformed series corresponds to the residuals of the model.

Deseasonalization

Goodness of fit coefficients: This table displays the goodness of fit coefficients.

Estimates of the parameters of the model: This table displays the parameters of the model. The intercept is equal to the mean of the series before transformation.

Series before and after transformation: This table displays the series before and after transformation. The transformed series corresponds to the residuals of the model.

Example

A tutorial explaining how to transform time series is available on the XLSTAT Help Center. To consult the tutorial, please go to:

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

References

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.

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)

Smoothing

Use this tool to smooth a time series and make predictions, using moving averages, exponential smoothing, Fourier smoothing, Holt or Holt-Winter's methods.

In this section:

Description

Several smoothing methods are available. We define by Y_t, (t=1,…,n), the time series of interest, by P_tY_{t+h} the predictor of Y_{t+h} with minimum mean square error, and ϵ_t follows a $N(0,1)$ white noise. The smoothing methods are described by the following equations:

Simple exponential smoothing

This model is sometimes referred to as Brown's Simple Exponential Smoothing, or the exponentially weighted moving average model. The equations of the model write:

$$\begin{cases} Y_t = \mu_t + \epsilon_t & \\ P_tY_{t+h} = \mu_t & h = 1, 2, ... \\ S_tY_{t+h} = \alpha Y_t + (1-\alpha)S_{t-1} & 0 < \alpha < 2 \\ \hat{Y}_{t+h} = P_tY_{t+h} = S_t, & h = 1, 2, ... \end{cases}$$

The region for α corresponds to the additivity and invertibility conditions.

Exponential smoothing is useful when one needs to model a value by simply taking into account past observations. It is called "exponential" because the weight of past observations decreases exponentially. This method it is not very satisfactory in terms of prediction, as the predictions are constant after n+1.

Double exponential smoothing

This model is sometimes referred to as Brown's Linear Exponential Smoothing or Brown's Double Exponential Smoothing. It allows to take into account a trend that varies with time. The predictions take into account the trend as it is for the last observed data. The equations of the model write:

$$\begin{cases} Y_t = \mu_t + \beta_t t + \epsilon_t & \\ P_t Y_{t+h} = \mu_t + \beta_t t & h = 1, 2, ... \\ S_t = \alpha Y_t + (1-\alpha) S_{t-1} & 0 < \alpha < 2 \\ T_t = \alpha S_t + (1-\alpha) S_{t-1} & \\ \hat{Y}_{t+h} = P_t Y_{t+h} = (2 + \frac{\alpha h}{1-\alpha}) S_t - (1 + \frac{\alpha h}{1-\alpha}) T_t, & \alpha \neq 1, h = 1, 2, ... \\ \hat{Y}_{t+h} = P_t Y_{t+h} = Y_t, & \alpha = 0, h = 1, 2, ... \end{cases}$$

The region for α corresponds to additivity and invertibility.

Holt's linear exponential smoothing

This model is sometimes referred to as the Holt-Winters non seasonal algorithm. It allows to take into account a permanent component and a trend that varies with time. This models adapts itself quicker to the data compared with the double exponential smoothing. It involves a second parameter. The predictions for t>n take into account the permanent component and the trend component. The equations of the model write:

$$\begin{cases} Y_t = \mu_t + \beta_t t + \epsilon_t & \\ P_t Y_{t+h} = \mu_t + \beta_t t & h = 1, 2, ... \\ S_t = \alpha Y_t + (1-\alpha)(S_{t-1} + T_{t-1}) & 0 < \alpha < 2 \\ T_t = \beta (S_t - S_{t-1}) + (1-\beta) T_{t-1} & 0 < \beta < 4/\alpha - 2 \\ \hat{Y}_{t+h} = P_t Y_{t+h} = S_t + h T_t, & h = 1, 2, ... \end{cases}$$

The region for α and β corresponds to additivity and invertibility.

Holt-Winters seasonal additive model

This method allows to take into account a trend that varies with time and a seasonal component with a period p. The predictions take into account the trend and the seasonality. The model is called additive because the seasonality effect is stable and does not grow with time. The equations of the model write:

$$\begin{cases} Y_t = \mu_t + \beta_t t + s_p(t) + \epsilon_t & \\ P_t Y_{t+h} = \mu_t + \beta_t t + s_p(t) & h = 1, 2, ... \\ S_t = \alpha (Y_t - S_{t-p} + (1-\alpha)(S_{t-1} + T_{t-1}) & \\ T_t = \beta (S_t - S_{t-1}) + (1-\beta) T_{t-1} & \\ D_t = \gamma (Y_t - S_t) + (1-\gamma) D_{t-p} & \\ \hat{Y}_{t+h} = P_t Y_{t+h} = S_t + h T_t + D_{t-p+h}, & h = 1, 2, ... \end{cases}$$

For the definition of the additive-invertible region please refer to Archibald (1990).

Holt-Winters seasonal multiplicative model

This method allows to take into account a trend that varies with time and a seasonal component with a period p. The predictions take into account the trend and the seasonality. The model is called multiplicative because the seasonality effect varies with time. The more the discrepancies between the observations are high, the more the seasonal component increases. The equations of the model write:

$$\begin{cases} Y_t = (\mu_t + \beta_t t)s_p(t) + \epsilon_t \\ P_t Y_{t+h} = (\mu_t + \beta_t t)s_p(t) & h = 1, 2, ... \\ S_t = \alpha(Y_t / S_{t-p}) + (1-\alpha)(S_{t-1} + T_{t-1}) \\ T_t = \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1} \\ D_t = \gamma(Y_t / S_t) + (1-\gamma)D_{t-p} \\ \hat{Y}_{t+h} = P_t Y_{t+h} = (S_t + hT_t)D_{t-p+h}, & h = 1, 2, ... \end{cases}$$

For the definition of the additive-invertible region please refer to Archibald (1990).

Note 1: for all the above models, XLSTAT estimates the values of the parameters that minimize the mean square error (MSE). However, it is also possible to maximize the likelihood, as, apart from the Holt-Winters multiplicative model, it is possible to write these models as ARIMA models. For example, the simple exponential smoothing is equivalent to an ARIMA(0,1,1) model, the double exponential smoothing is equivalent to an ARIMA(0,2,2) model, and the Holt-Winters additive model is equivalent to an ARIMA (0,1,p+1)(0,1,0). If you prefer to maximize the likelihood, we advise you to use the ARIMA procedure of XLSTAT.

Note 2: for all the above models, initial values for S, T and D, are required. XLSTAT offers several options, including backcasting to set these values. When backcasting is selected, the algorithm reverses the series, starts with simple initial values corresponding to the Y(x) option, then computes estimates and uses these estimates as initial values. The values corresponding to the various options for each method are described hereunder:

Simple exponential smoothing:

$$\begin{cases} Y(1): & S_1 = Y_1 \\ Mean(6): & S_1 = \sum_{i=1}^{6} Y_i / 6 \\ Backcasting \\ Optimized \end{cases}$$

Double exponential smoothing:

$$\begin{cases} Y(1): & S_1 = Y_1, \quad T_1 = Y_1 \\ Mean(6): & S_1 = \sum_{i=1}^{6} Y_i / 6, \quad T_1 = S_1 \\ Backcasting \end{cases}$$

Holt's linear exponential smoothing:

$$\begin{cases} 0: & S_1 = 0 \\ Backcasting & \end{cases}$$

Holt-Winters seasonal additive model:

$$\begin{cases} Y(1+p): & S_{1+p} = \sum_{i=}^{p} Y_i/p, \\ & T_{1+p} = 0 \\ & D_t = Y_t - (Y_1 + T_{1+p}(i-1)) \;\; i = 1, ...p \\ Backcasting & \end{cases}$$

Holt Winters seasonal multiplicative model:

$$\begin{cases} Y(1+p): & S_{1+p} = \sum_{i=}^{p} Y_i/p, \\ & T_{1+p} = 0 \\ & D_t = Y_t/(Y_1 + T_{1+p}(i-1)) \;\; i = 1, ...p \\ Backcasting & \end{cases}$$

Moving average

This model is a simple way to take into account past and optionally future observations to predict values. It works as a filter that is able to remove noise. While with the smoothing methods defined below, an observation influences all future predictions (even if the decay is exponential), in the case of the moving average the memory is limited to q. If the constant l is set to zero, the prediction depends on the past q values and on the current value, and if l is set to one, it also depends on the next q values. Moving averages are often used as filters, and not as way to do accurate predictions. However XLSTAT enables you to do predictions based on the moving average model that writes:

$$\begin{cases} Y_t = \mu_t + \epsilon_t \\ \hat{\mu}_t = \frac{\sum_{i=q}^{ql} w_i Y_{t+i}}{\sum_{i=q}^{ql} w_i} \end{cases}$$

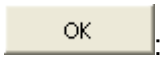
where l is a constant, which, when set to zero, allows the prediction to depend on the q previous values and on the current value. If l is set to one, the prediction also depends on the q next values. The $w_i (i = 1, \ldots, q)$ are the weights. Weights can be either constant, fixed by the user, or based on existing optimal weights for a given application. XLSTAT allows to use the Spencer 15-points model that passes polynomials of degree 3 without distortion.

Fourier smoothing

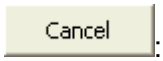
The concept of the Fourier smoothing is to transform a time series into its Fourier coordinates, then remove part of the higher frequencies, and then transform the coordinates back to a signal. This new signal is a smoothed series.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Times series: Select the data that correspond to the time series. If a header is available on the first row, make sure you activate the "Series labels" option.

Date data: Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format. If this option is not activated, XLSTAT creates its own time variable ranging from 1 to the number of data.

- **Check intervals**: Activate this option so that XLSTAT checks that the spacing between the date data is regular.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Series labels: Activate this option if the first row of the selected series includes a header.

Model: Select the smoothing model you want to use (see description for more information on the various models).

Options tab:

Method: Select the method for the selected model (see description for more information on the various models).

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 500.
- **Convergence**: Enter the maximum value of the evolution in the convergence criterion from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.00001.

Confidence interval (%): The value you enter (between 1 and 99) is used to determine the confidence intervals for the predicted values. Confidence intervals are automatically displayed on the charts.

S1: Choose an estimation method for the initial values. See the description for more information on that topic.

Depending on the model type, and on the method you have chosen, different options are available in the dialog box. In the description section, you can find information on the various models and on the corresponding parameters.

In the case of exponential or Holt-Winters models, you can decide to set the parameters to a given value, or to optimize them. In the case of the Holt- Winters seasonal models, you need to enter the value of the **period**.

In the case of the Fourier smoothing, you need to enter to the proportion **p** of the spectrum that needs to be kept after the high frequencies are removed.

For the moving average model, you need to specify the number **q** of time steps that must be taken into account to compute the predicted value. You can decided to only consider the previous q steps (the left part) of the series.

Validation tab:

Validation: Activate this option to use some data for the validation of the model.

Time steps: Enter the number the number of data at the end of the series that need to be used for the validation.

Prediction tab:

Prediction: Activate this option to use the model to do some forecasting.

Time steps: Enter the number the number of time steps for which you want XLSTAT to compute a forecast.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

- **Check for each series separately**: Check this option to remove observations with missing data series by series.
- **Across all series**: Check this option to remove observations with missing data for all series.

Replace by the average of the previous and next values: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

Ignore missing data: Activate this option to ignore missing data.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics of the selected series.

Goodness of fit coefficients: Activate this option to display the goodness of fit statistics.

Model parameters: Activate this option to display the table of the model parameters.

Predictions and residuals: Activate this option to display the table of the predictions and the residuals.

Charts tab:

Display charts: Activate this option to display the charts of the series before and after smoothing, as well as the bar chart of the residuals.

Results

Goodness of fit coefficients: This table displays the goodness of fit coefficients which include the number of degrees of freedom (DF), the DDL, the sum of squares of errors (SSE) the mean square of errors (MSE), the root of the MSE (RMSE), the mean absolute percentage error (MAPE), the mean percentage error (MPE) the mean absolute error (MAE) and the coefficient of determination (R^2). Note: all these statistics are computed for the observations involved in the estimation of the model only; the validation data are not taken into account.

Model parameters: This table displays the estimates of the parameters, and, if available, the standard error of the estimates. Note: to S1 corresponds the first computed value of the S series, and to T1 corresponds the first computed value of the series T. See the description for more information.

Series before and after smoothing: This table displays the series before and after smoothing. If some predictions have been computed (t>n), and if the confidence intervals option has been activated, the confidence intervals are computed for the predictions.

Charts: The first chart displays the data, the model, and the predictions (validation + prediction values) as well as the confidence intervals. The second chart corresponds to the bar chart of the residuals.

Example

A tutorial explaining how to do forecasting with the Holt-Winters method is available on the XLSTAT Help Center. To consult the tutorial, please go to:

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

References

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 Research 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

Use this tool to fit an ARMA (Autoregressive Moving Average), an ARIMA (Autoregressive Integrated Moving Average) a SARIMA (Seasonal Autoregressive Integrated Moving Average) model or a SARIMAX model (with exaplanatory variables), and to compute forecasts using the model which parameters are either known or to be estimated.

In this section:

Description

The models of the ARIMA family allow to represent in a synthetic way phenomena that vary with time, and to predict future values with a confidence interval around the predictions.

The mathematical writing of the ARIMA models differs from one author to the other. The differences concern most of the time the sign of the coefficients. XLSTAT is using the most commonly found writing, used by most software.

If we define by $\{X_t\}$ a series with mean μ, then if the series is supposed to follow an ARIMA $(p,d,q)(P,D,Q)^s$ model, we can write:

$$\begin{cases} Y_t = (1-B)^d(1-B^s)^D X_t - \mu \\ \phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t,\ Z_t \sim N(0,\sigma^2) \end{cases}$$

with

$$\begin{cases} \phi(z) = 1 - \sum_{i=1}^{p} \phi_i z^i, & \Phi(z) = 1 - \sum_{i=1}^{P} \Phi_i z^i \\ \theta(z) = 1 + \sum_{i=1}^{q} \theta_i z^i, & \Theta(z) = 1 + \sum_{i=1}^{Q} \Theta_i z^i \end{cases}$$

p is the order of the autoregressive part of the model.

q is the order of the moving average part of the model.

d is the differencing order of the model.

D is the differencing order of the seasonal part of the model.

s is the period of the model (for example 12 if the data are monthly data, and if one noticed a yearly periodicity in the data).

P is the order of the autoregressive seasonal part of the model.

Q is the order of the moving average seasonal part of the model.

Remark 1: the $\{Y_t\}$ process is causal if and only if for any z such that $|z| \leq 1$, $f(z) \neq 0$ and $q(z) \neq 0$.

Remark 2: if D=0, the model is an ARIMA(p,d,q) model. In that case, P, Q and s are considered as null.

Remark 3: if d=0 and D=0, the model simplifies to an ARMA(p,q) model.

Remark 4: if d=0, D=0 and q=0, the model simplifies to an AR(p) model.

Remark 5: if d=0, D=0 and p=0, the model simplifies to an MA(q) model.

Explanatory variables

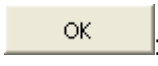
XLSTAT allows you to take into account explanatory variables through a linear model. Three different approaches are possible:

1. OLS: A linear regression model is fitted using the classical linear regression approach, then the residuals are modeled using an (S)ARIMA model.
2. CO-LS: If d or D and s are not zero, the data (including the explanatory variables) are differenced, then the corresponding ARMA model is fitted at the same time as the linear model coefficients using the Cochrane and Orcutt (1949) approach.
3. GLS: A linear regression model is fitted, then the residuals are modeled using an (S)ARIMA model, then we loop back to the regression step, in order to improve the likelihood of the model by changing the regression coefficients using a Newton-Raphson approach.

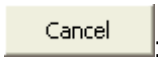
Note: if no differencing is requested (d=0 and D=0), and if there are no explanatory variables in the model, the constant of the model is estimated using CO-LS.

Dialog box

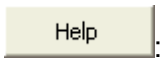
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Times series: Select the data that correspond to the time series. If a header is available on the first row, make sure you activate the "Series labels" option.

Center: Activate this option to center the data after the differencing.

Variance: Activate this option to set the value of the variance of the errors.

Date data: Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format. If this option is not activated, XLSTAT creates its own time variable ranging from 1 to the number of data.

- **Check intervals**: Activate this option so that XLSTAT checks that the spacing between the date data is regular.

X / Explanatory variables: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected must be of the numerical type. If a variable header has been selected, check that the "Variable labels" option has been activated.

- **Mode**: Choose the way you want to take into account the explanatory variables (the three modes OLS, CO-LS, GLS, are described in the description section).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Series labels: Activate this option if the first row of the selected series includes a header.

Model parameters: Enter orders of the model:

- **p**: Enter the order of the autoregressive part of the model. For example, enter 1 for an AR(1) model or for an ARMA(1,2) model.
- **d**: Enter the differencing order of the model. For example, enter 1 for an ARIMA(0,1,2) model.
- **q**: Enter the order of the moving average part of the model. For example, enter 2 for a MA(2) model or for an ARIMA(1,1,2) model.
- **P**: Enter the order of the autoregressive seasonal part of the model. For example, enter 1 for an ARIMA(1,1,0)(1,1,0)[12] model. You can modify this value only if $D \neq 0$. If $D = 0$, XLSTAT considers that $P = 0$.
- **D**: Enter the differencing order for the seasonal part of the model. For example, enter 1 for an ARIMA(0,1,1)(0,1,1)[12] model.
- **Q**: Enter the order of the moving average seasonal part of the model. For example, enter 1 for an ARIMA(0,1,1)(0,1,1)[12] model. You can modify this value only if $D \neq 0$. If $D = 0$, XLSTAT considers that $Q = 0$.
- **s**: Enter the period of the model. You can modify this value only if $D \neq 0$. If $D = 0$, XLSTAT considers that $s = 0$.

Options tab:

Preliminary estimation: Activate this option if you want to use a preliminary estimation method. This option is available only if D=0.

- **Yule-Walker**: Activate this option to estimate the coefficients of the autoregressive AR(p) model using the Yule-Walker algorithm.
- **Burg**: Activate this option to estimate the coefficients of the autoregressive AR(p) model using the Burg's algorithm.
- **Innovations**: Activate this option to estimate the coefficients of the moving average MA(q) model using the Innovations algorithm.
- **Hannan-Rissanen**: Activate this option to estimate the coefficients of the ARMA(p,q) model using the Hannan-Rissanen algorithm.
- **m/Automatic**: If you choose to use the Innovations or the Hannan-Rissanen algorithm, you need to either enter the m value corresponding to the algorithm or to let XLSTAT determine automatically (select Automatic) what is an appropriate value for m.

Initial coefficients: Activate this option to select the initial values of the coefficients of the model.

- **Phi**: Select here the value of the coefficients corresponding to the autoregressive part of the model (including the seasonal part). The number of values to select is equal to p+P.
- **Theta**: Select here the value of the coefficients corresponding to the moving average part of the model (including the seasonal part). The number of values to select is equal to q+Q.

Optimize: Activate this option to estimate the coefficients using one of the two available methods:

- **Likelihood**: Activate this option pour maximize the likelihood of the parameters knowing the data.
- **Least squares**: Activate this option to minimize the sum of squares of the residuals.

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 500.
- **Convergence**: Enter the maximum value of the evolution in the convergence criterion from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.00001.

Find the best model: Activate this option to explore several combines of orders. If you activate this option, the minimum order is the one given in the "General" tab, and the maximum orders need to be defined using the following options:

- **Max(p)**: Enter the maximum value of p to explore.
- **Max(q)**: Enter the maximum value of q to explore.
- **Max(P)**: Enter the maximum value of P to explore.
- **Max(Q)**: Enter the maximum value of Q to explore.
- **AICC**: Activate this option to use the AICC (Akaike Information Criterion Corrected) to identify the best model.
- **SBC**: Activate this option to use the SBC (Schwarz's Bayesian Criterion) to identify the best model.

Validation tab:

Validation: Activate this option to use some data for the validation of the model.

Time steps: Enter the number the number of data at the end of the series that need to be used for the validation.

Prediction tab:

Prediction: Activate this option to use the model to do some forecasting.

Time steps: Enter the number the number of time steps for which you want XLSTAT to compute a forecast.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

- **Check for each series separately**: Check this option to remove observations with missing data series by series.
- **Across all series**: Check this option to remove observations with missing data for all series.

Replace by the average of the previous and next values: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics of the selected series.

Goodness of fit coefficients: Activate this option to display the goodness of fit statistics.

Model parameters: Activate this option to display the table of the model parameters.

Predictions and residuals: Activate this option to display the table of the predictions and the residuals.

Confidence interval (%): The value you enter (between 1 and 99) is used to determine the confidence intervals for the predicted values. Confidence intervals are automatically displayed on the charts.

Charts tab:

Display charts: Activate this option to display the chart that shows the input data together with the model predictions, as well the bar chart of the residuals.

Results

Summary statistics: This table displays for the selected variables, the number of observations, the number of missing values, the number of non- missing values, the mean and the standard deviation (unbiased).

If a preliminary estimation and an optimization have been requested the results for the preliminary estimation are first displayed followed by the results after the optimization. If initial coefficients have been entered the results corresponding to these coefficients are displayed first.

Goodness of fit coefficients:

- **Observations**: The number of data used for the fitting of the model.
- **SSE**: Sum of Squares of Errors. This statistic is minimized if the "Least Squares" option has been selected for the optimization.
- **MAPE**: The Mean Absolute Percentage Error measures the quality of the fit, while removing the scale effect and not relatively penalizing bigger errors.
- **WN variance**: The white noise variance is equal to the SSE divided by N. In some software, this statistic is named sigma2 (sigma-square).
- **WN variance estimate**: This statistic is usually equal to the previous. In the case of a preliminary estimation using the Yule-Walker or Burg's algorithms, a slightly different estimate is displayed.
- **-2Log(Like.)**: This statistic is minimized if the "Likelihood" option has been selected for the optimization. It is equal to 2 times the natural logarithm of the likelihood.
- **FPE**: Akaike's Final Prediction Error. This criterion is adapted to autoregressive models.
- **AIC**: The Akaike Information Criterion.
- **AICC**: This criterion has been suggested by Brockwell (Akaike Information Criterion Corrected).
- **SBC**: Schwarz's Bayesian Criterion.

Model parameters:

The first table of parameters gives the coefficients of the linear model fitted to the data (a constant if no explanatory variable was selected).

The next table gives the estimator for each coefficient of each polynomial, as well as the standard deviation obtained either directly from the estimation method (preliminary estimation), or from the Fisher's information matrix (Hessian). The asymptotical standard deviations are also computed. For each coefficient and each standard deviation, a confidence interval is displayed. The coefficients are identified as follows:

AR(i): that corresponds to the order i coefficient of the f(z) polynomial.

SAR(i): coefficient that corresponds to the order i coefficient of the F(z) polynomial.

MA(i): coefficient that corresponds to the order i coefficient of the q(z) polynomial.

SMA(i): coefficient that corresponds to the order i coefficient of the Q(z) polynomial.

Data, Predictions and Residuals: This table displays the data, the corresponding predictions computed with the model, and the residuals. If the user requested it, predictions are computed for the validation data and forecasts for future values. Standard deviations and confidence intervals are computed for validation predictions and forecasts.

Charts: Two charts are displayed. The first chart displays the data, the corresponding values predicted by the model, and the predictions corresponding to the values for the validation and/or prediction time steps. The second chart corresponds to the bar chart of residuals.

Example

A tutorial explaining how to do fit an ARIMA model and to use the model to do forecasting is available on XLSTAT Help Center. To consult the tutorial, please go to:

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

References

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.

Spectral analysis

Use this tool to transform a time series into its coordinates in the space of frequencies, and then to analyze its characteristics in this space.

In this section:

Description

This tool allows to transform a time series into it's coordinates in the space of frequencies, and then to analyze its characteristics in this space. From the coordinates we can extract the magnitude, the phase, build representations such as the periodogram, the spectral density, and test if the series is stationary. By looking at the spectral density, we can identify seasonal components, and decide to which extent we should filter noise. Spectral analysis is a very general method used in a variety of domains.

The spectral representation of a time series $\{X_t\}$, $(t = 1, \ldots, n)$, decomposes $\{X_t\}$ into a sum of sinusoidal components with uncorrelated random coefficients. From there we can obtain decomposition the autocovariance and autocorrelation functions into sinusoids.

The spectral density corresponds to the transform of a continuous time series. However, we usually have only access to a limited number of equally spaced data, and therefore, we need to obtain first the discrete Fourier coordinates (cosine and sine transforms), and then the periodogram. From the periodogram, using a smoothing function, we can obtain a spectral density estimate which is a better estimator of the spectrum.

Using fast and powerful methods, XLSTAT automatically computes the Fourier cosine and sine transforms of $\{X_t\}$, for each Fourier frequency, and then the various functions that derive from these transforms.

With n being the sample size, and $[i]$ being the largest integer less than or equal to i, the Fourier frequencies write:

$$\omega_k = \frac{2\pi k}{n}, k = -[\frac{n-1}{2}], \ldots [\frac{n}{2}]$$

The Fourier cosine and sine coefficients write:

$$a_k = \frac{2}{n}\sum_{t=1}^{n} X_t \cos(\omega_k(t-1))$$

$$b_k = \frac{2}{n}\sum_{t=1}^{n} X_t \sin(\omega_k(t-1))$$

The periodogram writes:

$$I_k = \frac{n}{2}\sum_{t=1}^{n}(a_k^2 + b_k^2)$$

The spectral density estimate (or discrete spectral average estimator) of the time series $\{X_t\}$ writes:

$$\hat{f}_k = \sum_{i=-p}^{p} w_i J_{k+i}$$

with

$$\begin{cases} J_{k+i} = I_{k+i}, & 0 \leq k+i \leq n \\ J_{k+i} = I_{-(k+i)}, & k+i < 0 \\ J_{k+i} = I_{n-(k+i)}, & k+i > n \end{cases}$$

where p, the bandwidth, and w_i, the weights, are either fixed by the user, or determined by the choice of a kernel. XLSTAT suggests the use of the following kernels:

If we define , $p = c.q^e$, $q = [n/2]+1$, and $\lambda_i = i/p$

Bartlett:

$$\begin{cases} c = 1/2, & e = 1/3 \\ w_i = 1 - |\lambda_i| & \text{if } |\lambda_i| \leq 1 \\ w_i = 0 & \text{otherwise} \end{cases}$$

Parzen:

$$\begin{cases} c = 1, & e = 1/5 \\ w_i = 1 - 6|\lambda_i|^2 + 6|\lambda_i|^3 & \text{if } |\lambda_i| \leq 0.5 \\ w_i = 2(1-|\lambda_i|)^3 & \text{if } 0.5 \leq |\lambda_i| \leq 1 \\ w_i = 0 & \text{otherwise} \end{cases}$$

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:

$$\begin{cases} c = 2/3, & e = 1/5 \\ w_i = (1+\cos(\pi\lambda_i))/2 & \text{if } |\lambda_i| \leq 1 \\ w_i = 0 & \text{otherwise} \end{cases}$$

Truncated:

$$\begin{cases} c = 1/4, & e = 1/5 \\ w_i = 1 & \text{if } |\lambda_i| \leq 1 \\ w_i = 0 & \text{otherwise} \end{cases}$$

Note: the bandwidth p is a function of n, the size of the sample. The weights w_i must be positive and must sum to one. If they don't, XLSTAT automatically rescales them.

If a second time series $\{Y_t\}$ is available, several additional functions can be computed to estimate the cross-spectrum:

The real part of the cross-periodogram of $\{X_t\}$ and $\{Y_t\}$ writes:

$$Real_k = \frac{n}{2}\sum_{t=1}^{n}(a_{X,k}a_{Y,k} + b_{X,k}b_{Y,k})$$

The imaginary part of the cross-periodogram of $\{X_t\}$ and $\{Y_t\}$ writes:

$$Imag_k = \frac{n}{2}\sum_{t=1}^{n}(a_{X,k}b_{Y,k} - b_{X,k}a_{Y,k})$$

The cospectrum estimate (real part of the cross-spectrum) of the time series $\{X_t\}$ and $\{Y_t\}$ writes:

$$C_k = \sum_{i=-p}^{p} w_i R_{k+i}$$

with

$$\begin{cases} R_{k+i} = Real_{k+i}, & 0 \leq k+i \leq n \\ R_{k+i} = Real_{-(k+i)}, & k+i < 0 \\ R_{k+i} = Real_{n-(k+i)}, & k+i > n \end{cases}$$

The quadrature spectrum (imaginary part of the cross-periodogram) estimate of the time series $\{X_t\}$ and $\{Y_t\}$ writes:

$$Q_k = \sum_{i=-p}^{p} w_i H_{k+i}$$

with

$$\begin{cases} H_{k+i} = Imag_{k+i}, & 0 \leq k+i \leq n \\ H_{k+i} = Imag_{-(k+i)}, & k+i < 0 \\ H_{k+i} = Imag_{n-(k+i)}, & k+i > n \end{cases}$$

The phase of the cross-spectrum of $\{X_t\}$ and $\{Y_t\}$ writes:

$$\Phi_k = \arctan(Q_k / C_k)$$

The amplitude of the cross-spectrum of $\{X_t\}$ and $\{Y_t\}$ writes:

$$A_k = \sqrt{C_k^2 + Q_k^2}$$

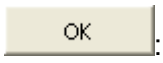
The squared coherency estimate between the $\{X_t\}$ and $\{Y_t\}$ series writes:

$$K_k = \frac{A_k^2}{\hat{f}_{X,k}\hat{f}_{Y,k}}$$

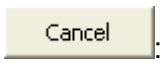
White noise tests: XLSTAT optionally displays two test statistics and the corresponding p-values for white noise: Fisher's Kappa and Bartlett's Kolmogorov-Smirnov statistic.

Dialog box

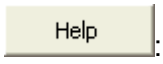
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Times series: Select the data that correspond to the time series for which you want to compute the various spectral functions.

Date data: Activate this option if you want to select date or time data. These data must be available either in the Excel date/time formats or in a numerical format. If this option is not activated, XLSTAT creates its own time variable ranging from 1 to the number of data.

- **Check intervals**: Activate this option so that XLSTAT checks that the spacing between the date data is regular.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Series labels: Activate this option if the first row of the selected series includes a header.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Replace by the average of the previous and next values: Activate this option to estimate the missing data by the mean of the first preceding non missing value and of the first next non missing value.

Outputs (1) tab:

White noise tests: Activate this option if you want to display the results of the white noise tests.

Cosine part: Activate this option if you want to display the Fourier cosine coefficients.

Sine part: Activate this option if you want to display the Fourier sine coefficients.

Amplitude: Activate this option if you want to display the amplitude of the spectrum.

Phase: Activate this option if you want to display the phase of the spectrum.

Spectral density: Activate this option if you want to display the estimate of spectral density.

- **Kernel weighting**: Select the type of kernel. The kernel functions are described in the description section.
- **c**: Enter the value of the c parameter. This parameter is described in the description section.

- **e**: Enter the value of the e parameter. This parameter is described in the description section.
- **Fixed weighting**: Select on an Excel sheet the values of the fixed weights. The number of weights must be odd. Symmetric weights are recommended (Example: 1,2,3,2,1).

Outputs (2) tab:

Cross-spectrum: Activate this option to analyze the cross-spectra. The computations are only done if at least two series have been selected.

- **Real part**: Activate this option to display the real part of the cross-spectrum.
- **Imaginary part**: Activate this option to display the imaginary part of the cross-spectrum.
- **Cospectrum**: Activate this option to display the cospectrum estimate (real part of the cross-spectrum).
- **Quadrature spectrum**: Activate this option to display the quadrature estimate (real part of the cross-spectrum).
- **Squared coherency**: Activate this option to display the squared coherency.

Charts tab:

Periodogram: Activate this option to display the periodogram of the series.

Spectral density: Activate this option to display the chart of the spectral density.

Results

White noise tests: This table displays both the Fisher's Kappa Bartlett's Kolmogorov-Smirnov statistics and the corresponding p-values. If the p-values are lower than the significance level (typically 0.05), then you need to reject the assumption that the times series is just a white noise.

A table is displayed for each selected time series. It displays various columns:

Frequency: frequencies from 0 to π.

Period: in time units.

Cosine part: the cosine coefficients of the Fourier transform.

Sine part: the sine coefficients of the Fourier transform.

Phase: Phase of the spectrum.

Periodogram: value of the periodogram.

Spectral density: estimate of the spectral density.

Charts: XLSTAT displays the periodogram and the spectral density charts on both the frequency and period scales.

If two series or more have been selected, and if the cross-spectrum options have been selected, XLSTAT displays additional information:

Cross-spectrum analysis: This table displays the cross-spectrum information:

Frequency: frequencies from 0 to π.

Period: in time units.

Real part: the real part of the cross-spectrum.

Imaginary part: the imaginary part of the cross-periodogram.

Cospectrum: The cospectrum estimate (real part of the cross-spectrum).

Quadrature spectrum: The quadrature estimate (imaginary part of the cross-spectrum).

Amplitude: amplitude of the cross-spectrum.

Squared coherency: estimates of the squared coherency.

Charts: XLSTAT displays the amplitude of the estimate of the cross- spectrum on both the frequency and period scales.

Example

An example of Spectral analysis is available on the XLSTAT Help Center. To consult the tutorial, please go to:

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

References

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.

Fourier transformation

Use this tool to transform a time series or a signal to its Fourier coordinates, or to do the inverse transformation.

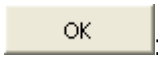
In this section:

Description

Use this tool to transform a time series or a signal to its Fourier coordinates, or to do the inverse transformation. While the Excel function is limited to powers of two for the length of the time series, XLSTAT is not restricted. Outputs optionally include the amplitude and the phase.

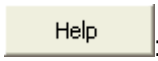
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.

: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Real part: Activate this option and then select the signal to transform, or the real part of the Fourier coordinates for an inverse transformation.

Imaginary part: Activate this option and then select the imaginary part of the Fourier coordinates for an inverse transformation.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selections (real part, imaginary part) includes a header.

Inverse transformation: Activate this option if you want to compute the inverse Fourier transform.

Amplitude: Activate this option if you want to compute and display the amplitude of the spectrum.

Phase: Activate this option if you want to compute and display the phase of the spectrum.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Results

Real part: This column contains the real part after the transform or the inverse transform.

Imaginary part: This column contains the real part after the transform or the inverse transform.

Amplitude: Amplitude of the spectrum.

Phase: Phase of the spectrum.

Example

References

Fuller W.A. (1996). Introduction to Statistical Time Series, Second Edition. John Wiley & Sons, New York.

Monte Carlo simulations

XLSTAT-Sim

XLSTAT-Sim is an easy to use and powerful solution to create and run simulation models.

In this section:

Introduction to XLSTAT-Sim

XLSTAT-Sim is a module that allows to build and compute simulation models, an innovative method for estimating variables, whose exact value is not known, but that can be estimated by means of repeated simulation of random variables that follow certain theoretical laws. Before running the model, you need to create the model, defining a series of input and output (or result) variables.

XLSTAT-Sim is a module that allows to build and compute simulation models, an innovative method for estimating variables, whose exact value is not known, but that can be estimated by means of repeated simulation of random variables that follow certain theoretical laws. Before running the model, you need to create the model, defining a series of input and output (or result) variables.

Simulation models

Simulation models allow to obtain information, such as mean or median, on variables that do not have an exact value, but for which we can know, assume or compute a distribution. If some "result" variables depend of these "distributed" variables by the way of known or assumed formulae, then the "result" variables will also have a distribution. XLSTAT-Sim allows you to define the distributions, and then obtain through simulations an empirical distribution of the input and output variables as well as the corresponding statistics.

Simulation models are used in many areas such as finance and insurance, medicine, oil and gas prospecting, accounting, or sales prediction.

Four elements are involved in the construction of a simulation model:

- Distribution s are associated to random variables. XLSTAT gives a choice of more than 20 distributions to describe the uncertainty on the values that a variable can take (see chapter Define a distribution for more details). For example, you can choose a triangular distribution if you have a quantity for which you know it can vary between two bounds, but with a value that is more likely (a mode). At each iteration of the computation of the simulation model, a random draw is performed in each distribution that has been defined.

- Scenario variables allow to include in the simulation model a quantity that is fixed in the model, except during the tornado analysis where it can vary between two bounds (see chapter Define a scenario variable for more details, and the section on tornado analysis below).

- Result variables correspond to outputs of the model. They depend either directly or indirectly, through one or more Excel formulae, on the random variables to which distributions have been associated and if available on the scenario variables. The goal of computing the simulation model is to obtain the distribution of the result variables (see chapter Define a result variable for more details).

- Statistics allow to track a given statistic a result variable. For example, we might want to monitor the standard deviation of a result variable (see chapter Define a statistic for more details).

A correct model should comprise at least one distribution and one result. Models can contain any number of these four elements.

A model can be limited to a single Excel sheet or can use a whole Excel folder.

Simulation models can take into account the dependencies between the input variables described by distributions. If you know that two variables are usually related such that the correlation coefficient between them is 0.4, then you want that, when you do simulations, the sampled values for both variables have the same property. This is possible in XLSTAT-Sim by entering in the Run dialog box the correlation or covariance matrix between some or all the input random variables used in the model.

Outputs

When you run the model, a series of results is displayed. While giving the critical statistics such are information on the distribution of the input and result variables, it also allows interpreting relationships between variables. Sensitivity analysis is also available if scenario variables have been included.

Descriptive statistics

The report that is generated after the simulation contains information on the distributions of the model. The user may choose from a range of descriptive statistics the most important indicators

that should be integrated into the report in order to easily interpret the results. A selection of charts is also available to graphically display the relationships.

Details and formulae relative to the descriptive statistics are available in the description section of the "Descriptive statistics " tool of XLSTAT.

Charts

The following charts are available to display information on the variables:

- **Box plots**: These univariate representations of quantitative data samples are sometimes called "box and whisker diagrams". It is a simple and quite complete representation since in the version provided by XLSTAT the minimum, 1-st quartile, median, mean and 3-rd quartile are displayed together with both limits (the ends of the "whiskers") beyond which values are considered anomalous. The mean is displayed with a red +, and a black line corresponds to the median. Limits are calculated as follows:

Lower limit: $L_{inf} = X(i)$ such that $\{X(i) - [Q1 - 1.5(Q3 - Q1)]\}$ is minimum and $X(i) = Q1 - 1.5(Q3 - Q1)$.

Upper limit: $L_{sup} = X(i)$ such that $\{X(i) - [Q3 + 1.5(Q3 - Q1)]\}$ is minimum and $X(i) = Q3 + 1.5(Q3 - Q1)$

Values that are outside the $]Q1 - 3(Q3 - Q1); Q3 + 3(Q3 - Q1)[$ interval are displayed with the * symbol; values that are in the $[Q1 - 3(Q3 - Q1); Q1 - 1.5(Q3 - Q1)]$ or the $[Q3 + 1.5(Q3 - Q1); Q3 + 3(Q3 - Q1)]$ intervals are displayed with the "o" symbol.

- **Scattergrams**: These univariate representations give an idea of the distribution and possible plurality of the modes of a sample. All points are represented together with the mean and the median.

- **P-P Charts (normal distribution)**: P-P charts (for Probability-Probability) are used to compare the empirical distribution function of a sample with that of a normal variable for the same mean and deviation. If the sample follows a normal distribution, the data will lie along the first bisector of the plan.

- **Q-Q Charts (normal distribution)**: Q-Q charts (for Quantile-Quantile) are used to compare the quantities of the sample with that of a normal variable for the same mean and deviation. If the sample follows a normal distribution, the data will lie along the first bisector of the plan.

Correlations

Once the computations are over, the simulation report may contain information on the correlations between the different variables included in the simulation model. Three different correlation coefficients are available:

Pearson correlation coefficient: This coefficient corresponds to the classical linear correlation coefficient. This coefficient is well suited for continuous data. Its value ranges from -1 to 1, and it measure the degree of linear correlation between two variables. Note: the squared Pearson

correlation coefficient gives an idea of how much of the variability of a variable is explained by the other variable. The p-values that are computed for each coefficient allow testing the null hypothesis that the coefficients are not significantly different from 0. However, one needs to be cautious when interpreting these results, as if two variables are independent, their correlation coefficient is zero, but the reciprocal is not true.

Spearman correlation coefficient (rho): This coefficient is based on the ranks of the observations and not on their value. This coefficient is adapted to ordinal data. As for the Pearson correlation, one can interpret this coefficient in terms of variability explained, but here we mean the variability of the ranks.

Kendall correlation coefficient (tau): As for the Spearman coefficient, it is well suited for ordinal variables as it is also based on ranks. However, this coefficient is conceptually very different. It can be interpreted in terms of probability: it is the difference between the probabilities that the variables vary in the same direction and the probabilities that the variables vary in the opposite direction. When the number of observations is lower than 50 and when there are no ties, XLSTAT gives the exact p-value. If not, an approximation is used. The latter is known as being reliable when there are more than 8 observations.

Sensitivity analysis

The sensitivity analysis displays information about the impact of the different input variables for one output variable. Based on the simulation results and on the correlation coefficient that has been chosen (see above), the correlations between the input random variables and the result variables are calculated and displayed in a declining order of impact on the result variable.

Tornado and spider analyses

Tornado and spider analyses are not based on the iterations of the simulation but on a point by point analysis of all the input variables (random variables with distributions and scenario variables).

During the tornado analysis, for each result variable, each input random variable and each scenario variable are studied one by one. We make their value vary between two bounds and record the value of the result variable in order to know how each random and scenario variable impacts the result variables. For a random variable, the values explored can either be around the median or around the default cell value, with bounds defined by percentiles or deviation. For a scenario variable, the analysis is performed between two bounds specified when defining the variables. The number of points is an option that can be modified by the user before running the simulation model.

The spider analysis does not only display the maximum and minimum change of the result variable, but also the value of the result variable for each data point of the random and scenario variables. This is useful to check if the dependence between distribution variables and result variables is monotonous or not.

Options

To display the options dialog box, click the  button of the "XLSTAT-SIM" toolbar. Use this dialog box to define the general options of the XLSTAT-SIM module.

General tab:

Model limited to: This option allows defining the size of the active simulation model. Limit if possible your model to a single Excel sheet. The following options are available:

- **Sheet**: Only the simulation functions in the active Excel sheet will be used in the simulation model. The other sheets are ignored.
- **Workbook**: All the simulation functions of the active workbook are included in the simulation model. This option allows using several Excel sheets for one model.

Sampling m ethod: This option allows choosing the method of sample generation. Two possibilities are available:

- **Classic**: The samples are generated using Monte Carlo simulations.
- **Latin h ypercubes**: The samples are generated using the Latin Hypercubes method. This method divides the distribution function of the variable into sections that have the same size and then generates equally sized samples within each section. This leads to a faster convergence of the simulation. You can enter the number of **sections.** Default value is 500.

Single step memory: Enter the maximum number of simulation steps that will be stored in the single step mode in order to calculate the statistics fields. When the limit is reached, the window moves forward (the first iteration is forgotten and the new one is stored). The default value is 500. This value can be larger, if necessary.

Number of iterations by step: Enter the value of the number of simulation iterations that are performed during one step. The default value is 1.

Format tab:

Use these options to set the format of the various model elements that are displayed on the Excel sheets:

- **Distributions**: You can define the color of the font and the color of the background of the cells where the definition of the input random variables and their corresponding distributions are stored.
- **Scenario variables**: You can define the color of the font and the color of the background of the cells where the scenario variables are stored.
- **Result variables**: You can define the color of the font and the color of the background of the cells where the result variables are stored.
- **Statistics**: You can define the color of the font and the color of the background of the cells where the statistics are stored.

Convergence tab:

Stop conditions: Activate this option to stop the simulation if the convergence criteria are reached.

- **Criterion**: Select the criterion that should be used for testing the convergence. There are three options available:
- **Mean**: The means of the monitored "result variables" (see below) of the simulation model will be used to check if the convergence conditions are met.
- **Standard deviation**: The standard deviation of the monitored "result variables" (see below) of the simulation model will be used to check if the convergence conditions are met.
- **Percentile**: The percentiles of the monitored "result variables" (see below) of the simulation model will be used to check if the convergence conditions are met. Choose the **Percentile** to be used. Default value is 90%.
- **Test frequency**: Enter the number of iterations to perform before the convergence criteria are checked again. Default value: 100.
- **Convergence**: Enter the value in % of the evolution of the convergence criteria from one check to the next, which, when reached, means that the algorithm has converged. Default value: 3%.
- **Confidence interval (%)**: Enter the size in % of the confidence interval that is computed around the selected criterion. The upper bound of the interval is compared to the convergence value defined above, in order to determine if the convergence is reached or not. Default value: 95%.
- **Monitored results**: Select which result variables of the simulation model should be monitored for the convergence. There are two options available:
- **All result variables**: All result variables of the simulation model will be monitored during the convergence test.
- **Activated result variables**: Only result variables that have their ConvActive parameter equal to 1 are monitored.

References tab:

Reference to Excel cells: Select the way references to names of variables to the simulation models are generated:

- **Absolute reference**: XLSTAT creates absolute references (for example $A4$) to the cell.
- **Relative reference**: XLSTAT creates absolute references (for example A4) to the cell.

Note: The absolute reference will not be changed if you copy and paste the XLSTAT_Sim formula, contrary to the relative reference.

Results tab:

Filter level for results: Select the level of details that will be displayed in the report. This controls for the descriptive statistics tables and the histograms of the different model elements:

- **All**: Details are displayed for all elements of the model.
- **Activated**: Details are only displayed for the elements that have a value of the Visible parameter set to 1.
- **None**: No detail will be displayed for the elements of the model.

Toolbar

XLSTAT-Sim has a dedicated toolbar "XLSTAT-Sim". The "XLSTAT-Sim" toolbar can be displayed by clicking the XLSTAT-Sim icon SIM in the XLSTAT toolbar.

Click this icon to define a new distribution (see Define a distribution for more details).

Click this icon to define a new scenario variable (see Define a scenario variable for more details).

Click this icon to define a new result (see Define a result variable for more details).

Click this icon to define a new statistic (see Define a statistic for more details).

Click this icon to reinitialize the simulation model and do a first simulation iteration.

Click this icon to do one simulation step.

Click this icon to start the simulation and display a report.

Click this icon to export the simulation model. All XLSTAT-Sim functions are transformed to comments. The formulae in the cells are stored as cell comments and the formulae are either replaced by the default value or by the formula linking to other cells in the case of XLSTAT_SimRes.

Click this icon to import the simulation model. All XLSTAT-Sim functions are extracted from cell comments and exported as formulae in the corresponding cells.

Click this icon to display the XLSTAT-Sim options dialog box.

Example

Examples showing how to build a simulation model are available on the XLSTAT Help Center at:

http://www.xlstat.com/demo-sim1.htm

http://www.xlstat.com/demo-sim2.htm

http://www.xlstat.com/demo-sim3.htm

http://www.xlstat.com/demo-sim4.htm

References

Vose, D. (2008). Risk Analysis – A Quantitative Guide, Third Edition, John Wiley & Sons, New York.

Define a distribution

Use this tool in a simulation model when there is uncertainty on the value of a variable (or quantity) that can be described with a distribution. The distribution will be associated with the currently selected cell.

In this section:

Description

This function is one of the essential elements of a simulation model. For a more detailed description on how a simulation model is constructed and calculated, please read the introduction on XLSTAT-Sim.

This tool allows to define the theoretical distribution function with known parameters that will be used to generate a sample of a given random variable. A wide choice of distribution functions is available.

To define the distribution that a given variable (physically, a cell on the Excel sheet) follows, you need to create a call to one of the XLSTAT_SimX functions or to use the dialog box that will generate for you the formula calling XLSTAT_SimX. X stands for the distribution (see the table below for additional details).

XLSTAT_SimX syntax:

XLSTAT_SimX(VarName, Param1, Param2, Param3, Param4, Param5, TruncMode, LowerBound, UpperBound, DefaultType, DefaultValue, Visible)

XLSTAT_SimX stands for one of the available distribution functions that are listed in the table below. A variable based on the corresponding distribution is defined. See the table below to see the available distributions.

VarName is a string giving the name of the variable for which the distribution is being defined. The name of the variable is used in the report to identify the variable.

Param1 is an optional input (default is 0) that gives the value of the first parameter of the distribution if relevant.

Param2 is an optional input (default is 0) that gives the value of the second parameter of the distribution if relevant.

Param3 is an optional input (default is 0) that gives the value of the third parameter of the distribution if relevant.

Param4 is an optional input (default is 0) that gives the value of the fourth parameter of the distribution if relevant.

Param5 is an optional input (default is 0) that gives the value of the fifth parameter of the distribution if relevant.

TruncMode is an optional integer that indicates if and how the distribution is truncated. A 0 (default value) corresponds to no truncation. 1 corresponds to truncating the distribution between two bounds that must then be specified. 2 corresponds to truncating between two percentiles that must then be specified.

TruncLower is an optional value that gives the lower bound of the truncation.

TruncUpper is an optional value that gives the upper bound of the truncation.

DefaultType is an optional integer that chooses the default value of the variable: 0 (default value) corresponds to the theoretical expected mean; 1 to the value given by the DefaultValue argument.

DefaultValue is an optional value giving the default value displayed in the cell before any simulation is performed. When no simulation process is ongoing, the default value will be displayed in the Excel cell as the result of the function.

Visible is an optional input that indicates if the details of this variable should be displayed in the simulation report. This option is only taken into account when the "Filter level for results" in the Options dialog box of XLSTAT-Sim is set to "Activated" (see the Format tab). 0 deactivates the display and 1 activates the display. Default value is 1.

Distribution	XLSTAT Name	Param1	Param2	Param3	Param4	Param5
Bernoulli	XLSTAT_SimBernoulli	p				
Beta	XLSTAT_SimBeta	alpha	beta			
Beta4	XLSTAT_SimBeta4	alpha	beta	c	d	
Binomial	XLSTAT_SimBinomial	n	p			
Chi square	XLSTAT_SimChiSqr	df				
Erlang	XLSTAT_SimErlang	k	gamma			
Exponential	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					
Logistic	XLSTAT_SimLogistic	mu	sigma			
Lognormal	XLSTAT_SimLognormal	mu	sigma			
Lognormal2	XLSTAT_SimLognormal2	m	s			
Negative Binomial (1)	XLSTAT_SimNegBinomial1	n	p			
Negative Binomial (2)	XLSTAT_SimNegBinomial2	k	p			
Normal	XLSTAT_SimNormal	mu	sigma			
Normal (Standard)	XLSTAT_SimNormalStd					
Pareto	XLSTAT_SimPareto	a	b			
Pert	XLSTAT_SimPert	a	m	b		
Poisson	XLSTAT_SimPoisson	Lambda				
Student	XLSTAT_SimStudent	df				
Trapezoidal	XLSTAT_SimTrapezoidal	a	b	c	d	
Triangular	XLSTAT_SimTriangular	a	m	b		
TriangularQ	XLSTAT_SimTriangularQ	a	m	b	q1	q2
Uniform	XLSTAT_SimUniform	a	b			
Uniform discrete	XLSTAT_SimUniformDisc	a	b			
Weibull (1)	XLSTAT_SimWeibull1	beta				
Weibull (2)	XLSTAT_SimWeibull2	beta	gamma			
Weibull (3)	XLSTAT_SimWeibull3	beta	gamma	mu		

Example:

=XLSTAT_SimNormal("Revenue Q1", 50000, 5000)

The function will associate to the cell where they are entered a normal distribution with mean 50000 and standard deviation 5000. The cell will show 50000 (the default value). If a report is generated afterwards the results corresponding to that cell will be identified by "Revenue Q1". The Param3, Param4 and Param5 are not entered because the Normal distribution has only two parameters. As the other parameters are not entered, they are set to their default value.

Determination of the parameters

In general, the choice of law and parameters of the law is guided by an empirical knowledge of the phenomenon, the results already available or working hypothesis.

To select the best suited law and the corresponding parameters you can use the Distribution fitting tool of XLSTAT. If you have a sample of data, by the help of this tool you can find the best parameters for a given distribution.

Random distributions available in XLSTAT-Sim

XLSTAT provides the following distributions:

- Arcsine (a): the density function of this distribution (which is a simplified version of the Beta type I distribution) is given by:

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

We have $E(X) = \alpha$ and $V(X) = \alpha(1-\alpha)/2$

- Bernoulli (p): the density function of this distribution is given by:

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

We have $E(X) = p$ and $V(X) = p(1-p)$

The Bernoulli, named after the Swiss mathematician Jacob Bernoulli (1654-1705), allows to describe binary phenomena where only events can occur with respective probabilities of p and $1-p$.

- Beta (a, b): the density function of this distribution (also called Beta type I) is given by:

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

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

We have $E(X) = \alpha/(\alpha+\beta)$ and $V(X) = \alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2]$

- Beta4 (α, β, c, d): the density function of this distribution is given by:

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

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

We have $E(X) = \frac{c+(c-d)\alpha}{(\alpha+\beta)}$ and $V(X) = \frac{(c-d)^2\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$

Pour the type I beta distribution, X takes values in the $[0,1]$ range. The beta4 distribution is obtained by a variable transformation such that the distribution is on a $[c,d]$ interval where c and d can take any value.

- Binomial (n, p) : the density function of this distribution is given by:

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

We have $E(X) = np$ and $V(X) = np(1-p)$

n is the number of trials, and p the probability of success. The binomial distribution is the distribution of the number of successes for n trials, given that the probability of success is p.

- Negative binomial type I (n, p): the density function of this distribution is given by:

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

We have $E(X) = n(1-p)/p$ and $V(X) = n(1-p)/p^2$

n is the number of successes, and p the probability of success. The negative binomial type I distribution is the distribution of the number x of unsuccessful trials necessary before obtaining n successes.

- Negative binomial type II (k, p): the density function of this distribution is given by:

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

We have $E(X) = kp$ and $V(X) = kp(p+1)$

The negative binomial type II distribution is used to represent discrete and highly heterogeneous phenomena. As k tends to infinity, the negative binomial type II distribution tends towards a Poisson distribution with $\lambda = kp$.

- $Khi^2(df)$: the density function of this distribution is given by:

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

We have $E(X) = df$ and $V(X) = 2df$

The Chi-square distribution corresponds to the distribution of the sum of df squared standard normal distributions. It is often used for testing hypotheses.

- Erlang (k, λ): the density function of this distribution is given by:

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

We have $E(X) = k/\lambda$ and $V(X) = k/\lambda^2$

k is the shape parameter and λ is the rate parameter.

This distribution, developed by the Danish scientist A. K. Erlang (1878-1929) when studying the telephone traffic, is more generally used in the study of queuing problems.

Note: When $k = 1$, this distribution is equivalent to the exponential distribution. The Gamma distribution with two parameters is a generalization of the Erlang distribution to the case where k is a real and not an integer (for the Gamma distribution the scale parameter $\beta = 1/\lambda$ is used).

- Exponential(λ): the density function of this distribution is given by:

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

We have $E(X) = 1/\lambda$ and $V(X) = 1/\lambda^2$

The exponential distribution is often used for studying lifetime in quality control.

- Fisher (df_1, df_2): the density function of this distribution is given by:

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

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

We have $E(X) = df_2/(df_2 - 2)$ if $df_2 > 2$, and $V(X) = \frac{2df_2^2(df_1+df_2-2)}{df_1(df_2-2)^2(df_2-4)}$ if $df_2 > 4$

Fisher's distribution, from the name of the biologist, geneticist and statistician Ronald Aylmer Fisher (1890-1962), corresponds to the ratio of two Chi-square distributions. It is often used for testing hypotheses.

- Fisher-Tippett (β, μ): the density function of this distribution is given by:

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

We have $E(X) = \mu + \beta\gamma$ and $V(X) = (\pi\beta)^2/6$ where γ is the Euler-Mascheroni constant.

The Fisher-Tippett distribution, also called the Log-Weibull or extreme value distribution, is used in the study of extreme phenomena. The Gumbel distribution is a special case of the Fisher-Tippett distribution where $\beta = 1$ and $\mu = 0$.

- Gamma (k, β, μ): the density of this distribution is given by:

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

We have $E(X) = \mu + k\beta$ and $V(X) = k\beta^2$

k is the shape parameter of the distribution and β the scale parameter.

- GEV (β, k, μ): the density function of this distribution is given by:

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

We have $E(X) = \mu + \frac{\beta}{k}\Gamma(1+k)$ and $V(X) = \left(\frac{\beta}{k}\right)^2 \left(\Gamma(1+2k) - \Gamma^2(1+k)\right)$

The GEV (Generalized Extreme Values) distribution is much used in hydrology for modeling flood phenomena. k lies typically between -0.6 and 0.6.

- Gumbel: the density function of this distribution is given by:

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

We have $E(X) = \gamma$ and $V(X) = \pi^2/6$ where γ is the Euler-Mascheroni constant (0.5772156649…).

The Gumbel distribution, named after Emil Julius Gumbel (1891-1966), is a special case of the Fisher-Tippett distribution with $\beta = 1$ and $\mu = 0$. It is used in the study of extreme phenomena such as precipitations, flooding and earthquakes.

- Logistic (μ, s): the density function of this distribution is given by:

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

We have $E(X) = \mu$ and $V(X) = (\pi s)^2/3$

- Lognormal (μ, σ): the density function of this distribution is given by:

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

We have $E(X) = exp(\mu + \sigma^2/2)$ and $V(X) = \left(exp(\sigma^2) - 1\right) exp(2\mu + \sigma^2)$

- Lognormal2 (m, s): the density function of this distribution is given by:

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

We have:

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

And:

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

This distribution is just a reparametrization of the Lognormal distribution.

- Normal (μ, σ) : the density function of this distribution is given by:

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

We have $E(X) = \mu$ and $V(X) = \sigma^2$

- Standard normal: the density function of this distribution is given by:

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

We have $E(X) = 0$ and $V(X) = 1$

This distribution is a special case of the normal distribution with $\mu = 0$ and $\sigma = 1$

- Pareto (a, b): the density function of this distribution is given by:

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

We have $E(X) = ab/(a-1)$ with $V(X) = \frac{ab^2}{((a-1)^2(a-2))}$

The Pareto distribution, named after the Italian economist Vilfredo Pareto (1848-1923), is also known as the Bradford distribution. This distribution was initially used to represent the distribution of wealth in society, with Pareto's principle that 80% of the wealth was owned by 20% of the population.

- PERT (a, m, b): the density function of this distribution is given by:

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

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

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

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

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

The PERT distribution is a special case of the beta4 distribution. It is defined by its definition interval [a, b] and m the most likely value (the mode). PERT is an acronym for *Program Evaluation and Review Technique*, a project management and planning methodology. The PERT methodology and distribution were developed during the project held by the US Navy and Lockheed between 1956 and 1960 to develop the Polaris missiles launched from submarines. The PERT distribution is useful to model the time that is likely to be spent by a team to finish a project. The simpler triangular distribution is similar to the PERT distribution in that it is also defined by an interval and a most likely value.

- Poisson (λ): the density function of this distribution is given by:

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

We have $E(X) = \lambda$ with $V(X) = \lambda$

Poisson's distribution, discovered by the mathematician and astronomer Siméon- Denis Poisson (1781-1840), pupil of Laplace, Lagrange and Legendre, is often used to study queuing phenomena.

- Student (df) : the density function of this distribution is given by:

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

We have $E(X) = 0$ if $df > 1$ with $V(X) = df/(df - 2)$ if $df > 2$

The English chemist and statistician William Sealy Gosset (1876-1937), used the nickname Student to publish his work, in order to preserve his anonymity (the Guinness brewery forbade its employees to publish following the publication of confidential information by another researcher). The Student's t distribution is the distribution of the mean of df variables standard normal variables. When $df = 1$, Student's distribution is a Cauchy distribution with the particularity of having neither expectation nor variance.

- Trapezoidal (a, b, c, d): the density function of this distribution is given by:

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

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

This distribution is useful to represent a phenomenon for which we know that it can take values between two extreme values (a and d), but that it is more likely to take values between two values (b and c) within that interval.

- Triangular (a, m, b): the density function of this distribution is given by:

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

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

- TriangularQ (q_1, m, q_2, p_1, p_2): the density function of this distribution is a reparametrization of the Triangular distribution. A first step requires estimating the a and b parameters of the triangular distribution, from the q_1 and q_2 quantiles to which percentages p_1 and p_2 correspond. Once this is done, the distribution functions can be computed using the triangular distribution functions.

- Uniform (a, b): the density function of this distribution is given by:

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

We have $E(X) = (a + b)/2$ with $V(X) = (b - a)^2/12$

The uniform (0,1) distribution is much used for simulations. As the cumulative distribution function of all the distributions is between 0 and 1, a sample taken in a Uniform (0,1) distribution is used to obtain random samples in all the distributions for which the inverse can be calculated.

- Uniform discrete (a, b): the density function of this distribution is given by:

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

We have $E(X) = (a + b)/2$ with $V(X) = [(b - a + 1)^2 - 1]/12$

The uniform discrete distribution corresponds to the case where the uniform distribution is restricted to integers.

- Weibull (β): the density function of this distribution is given by:

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

We have $E(X) = \Gamma(\frac{1}{\beta} + 1)$ with $V(X) = \Gamma(\frac{2}{\beta} + 1) - \Gamma^2(\frac{1}{\beta} + 1)$

β is the shape parameter for the Weibull distribution.

- Weibull (β, γ): the density function of this distribution is given by:

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

We have $E(X) = \gamma\Gamma\left(\frac{1}{\beta} + 1\right)$ with $V(X) = \gamma^2\left[\Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right)\right]$

β is the shape parameter of the distribution and γ the scale parameter. When $\beta = 1$, the Weibull distribution is an exponential distribution with parameter $1/\gamma$.

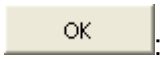
- Weibull (β, γ, μ): the density function of this distribution is given by:

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

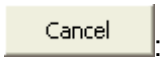
We have $E(X) = \mu + \gamma\Gamma\left(\frac{1}{\beta} + 1\right)$ with $V(X) = \gamma^2\left[\Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right)\right]$

The Weibull distribution, named after the Swede Ernst Hjalmar Waloddi Weibull (1887-1979), is much used in quality control and survival analysis. β is the shape parameter of the distribution and γ the scale parameter. When $\beta = 1$ and $\mu = 0$, the Weibull distribution is an exponential distribution with parameter $1/\gamma$.

Dialog box



: click this button to create the variable.



: click this button to close the dialog box without doing any modification.

: click this button to display help.


: click this button to reload the default options.


: click this button to delete the data selections.

General tab:

Variable name: Enter the name of the random variable or select a cell where the name is available. If you select a cell, an absolute reference (for example $A4$) or a relative reference (for

example A4) to the cell is created, depending on your choice in the XLSTAT options. (See the Options section for more details)

Distributions: Select the distribution that you want to use for the simulation. See the description section for more information on the available distributions.

Parameters: Enter the value of the parameters of the distribution you selected.

Truncation: Activate this option to truncate the distribution.

- **Absolute**: Select this option, if you want to enter the lower and upper bound of the truncation as absolute values.
- **Percentile**: Select this option, if you want to enter the lower and upper bound of the truncation as percentile values.
- **Lower bound**: Enter the value of the lower bound of the truncation.
- **Upper bound**: Enter the value of the upper bound of the truncation.

Options tab:

Default cell value: Choose the default value of the random variable. This value will be returned when no simulation model is running. The value may be defined by one of the following three methods:

- **Expected value**: This option selects the expected value of the distribution as the default cell value.
- **Fixed value**: Enter the default value.
- **Reference**: Choose a cell in the active Excel sheet that contains the default value.

Display results: Activate this option to display the detailed results for the random variable in the simulation report. This option is only active if you selected the "Activated" filter level in the simulation preferences. (See the Options section for more details).

Results

The result is function call to XLSTAT_SimX with the selected parameters. The following formula is generated in the active Excel cell:

= XLSTAT_SimX(VarName, Param1, Param2, Param3, Param4, Param5, TruncMode, LowerBound, UpperBound, DefaultType, DefaultValue, Visible)

The background color and the font color in the Excel cell are applied according to your choices in the XLSTAT-Sim options.

Define a scenario variable

Use this tool to define a variable which value varies between two known bounds during the tornado analysis.

In this section:

Description

This function allows to build a scenario variable that is used during the tornado analysis. For a more detailed description on how a simulation model is constructed, please read the introduction on XLSTAT-Sim.

A scenario variable is used for tornado analysis. This function gives you the possibility to define a scenario variable by letting XLSTAT know the bounds between which it varies. To define the scenario variable (physically, a cell on the Excel sheet), you need to create a call to the XLSTAT_SimSVar function or to use the dialog box that will generate for you the formula calling XLSTAT_SimSVar.

XLSTAT_SimSVar syntax

XLSTAT_SimSVar(SVarName, LowerBound, UpperBound, Type, Step, DefaultType, DefaultValue, Visible)

SVarName is a string that contains the name of the scenario variable. This can be a reference to a cell in the same Excel sheet. The name is used during the report to identify the cell.

LowerBound corresponds to the lower bound of the interval of possible values for the scenario variable.

UpperBound corresponds to the upper bound of the interval of possible values for the scenario variable.

Type is an integer that indicates the data type of the scenario variable. 1 stands for a continuous variable and 2 for a discrete variable. This input is optional with default value 1.

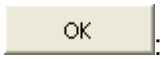
Step is a number that indicates in the case of a discrete variable the step size between two values to be examined during the tornado analysis. This input is optional with default value 1.

DefaultType is an optional integer that chooses the default value of the variable: 0 (default value) corresponds to the theoretical expected mean; 1 to the value given by the DefaultValue argument.

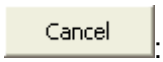
DefaultValue is a value that that corresponds to the default value of the scenario variable. The default value is returned as the result of this function.

Visible is an optional input that indicates if the details of this variable should be displayed in the simulation report. This option is only taken into account when the "Filter level for results" in the Options dialog box of XLSTAT-Sim is set to "Activated" (see the Format tab). 0 deactivates the display and 1 activates the display. Default value is 1.

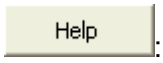
Dialog box



: click this button to create the variable.



: click this button to close the dialog box without doing any modification.



: click this button to display help.


: click this button to reload the default options.


: click this button to delete the data selections.

General tab:

Variable name: Enter the name of the scenario variable or select a cell where the name is available. If you select a cell, an absolute reference (for example $A4$) or a relative reference (for example A4) to the cell is created, depending on your choice in the XLSTAT options. (See the Options section for more details)

Lower bound: Enter the value of the lower bound or select a cell in the active Excel sheet that contains the value of the lower bound of the interval in which the scenario variable varies.

Upper bound: Enter the value of the upper bound or select a cell in the active Excel sheet that contains the value of the upper bound of the interval in which the scenario variable varies.

Data type:

- **Continuous**: Choose this option to define a continuous scenario variable that can take any value between the lower and upper bounds.
- **Discrete**: Choose this option to define a discrete scenario variable.
- **Step**: Enter the value of the step or select a cell in the active Excel sheet that contains the value of the step.

Options tab:

Default cell value: Choose the default value of the random variable. This value will be returned when no simulation model is running. The value may be defined by one of the following three methods:

- **Expected value**: This option returns the center of the interval as the default cell value.
- **Fixed value**: Enter the default value.
- **Reference**: Choose a cell in the active Excel sheet that contains the default value.

Display results: Activate this option to display the detailed results for the random variable in the simulation report. This option is only active if you selected the "Activated" filter level in the simulation preferences. (See the Options section for more details).

Results

The result is function call to XLSTAT_SimSVar with the selected parameters. The following formula is generated in the active Excel cell:

=XLSTAT_SimSVar(SVarName, LowerBound, UpperBound, Type, Step, DefaultType, DefaultValue, Visible)

The background color and the font color in the Excel cell are applied according to your choices in the XLSTAT-Sim options.

Define a result variable

Use this tool in a simulation model to define a result variable which calculation is the real aim of the simulation model.

In this section:

Description

This result variable is one of the two essential elements of a simulation model. For a more detailed description on how a simulation model is constructed and calculated, please read the introduction on XLSTAT-Sim.

Result variables can be used to define when a simulation process should stop during a run. If, in the XLSTAT-Sim Options dialog box, you asked that the "Activated result variables" are used the stop the simulations when, for example the mean has converged, then, if the ConvActiv parameter of the result variable is set to 1, the mean of the variable will used to determine if the simulation process has converged or not.

To define the result variable (physically, a cell on the Excel sheet), you need to create a call to the XLSTAT_SimRes function or to use the dialog box that will generate for you the formula calling XLSTAT_SimRes.

XLSTAT_SimRes syntax:

XLSTAT_SimRes (ResName, Formula, DefaultValue, ConvActiv, Visible)

ResName is a string that contains the name of the result variable or a reference to a cell where the name is located. The name is used during the report to identify the result variable.

Formula is a string that contains the formula that is used to calculate the results. The formula links directly or indirectly the random input variables and, if available the scenario variables, to the result variable. This corresponds to an Excel formula without the leading "=".

DefaultValue of type number is optional and contains the default value of the result variable. This value is not used in the computations.

ConvActiv is an integer that indicates if this result is checked during the convergence tests. This option is only active, if the "Activated result variables" convergence option is activated in the XLSTAT-Sim Options dialog box.

Visible is an optional input that indicates if the details of this variable should be displayed in the simulation report. This option is only taken into account when the "Filter level for results" in the

Options dialog box of XLSTAT-Sim is set to "Activated" (see the Format tab). 0 deactivates the display and 1 activates the display. Default value is 1.

Example:

=XLSTAT_SimRes("Forecast N+1", B3+B4-B5)

This function defines in the active cell a result variable called "Forecast N +1" calculated as the sum of cells B3 and B4 minus B5. The Visible parameter is not entered because it is only necessary when the "Filter level for the results" is set to "Activated" (see the Options dialog box) and because we want the result to be anyway visible.

Dialog box

OK: click this button to create the variable.

Cancel: click this button to close the dialog box without doing any modification.

Help: click this button to display help.

 : click this button to reload the default options.

 : click this button to delete the data selections.

General tab:

Variable name: Enter the name of the random variable or select a cell where the name is available. If you select a cell, it depends on the selection in the options, whether an absolute (for example $A4$) or a relative reference (for example A4) to the cell is created. (See the Options section for more details)

Use to monitor convergence: Activate this option to include this result variable in the result variables that are used to test for convergence. This option is only active, if you selected the "Activated results variables" option in the XLSTAT-Sim convergence options. ConvActiv should be 1 if you want the variable to be used to monitor the results. Default value is 1.

Display Results: Activate this option to display the detailed results for the result variable in the simulation report. This option is only active, if you selected the restricted filter level in the simulation preferences. (See the XLSTAT-Sim options for more details).

Results

A function call to XLSTAT_SimRes with the selected parameters and the following syntax will be generated in the active Excel cell:

=XLSTAT_SimRes (ResName, Formula, DefaultValue, ConvActiv, Visible)

The background color and the font color in the Excel cell are applied according to your choices in the XLSTAT-Sim options.

Define a statistic

Use this tool in a simulation model to define a statistic based on a variable of the simulation model. The statistic is updated after each iteration of the simulation process. Results relative to the defined statistics are available in the simulation report. A wide choice of statistics is available.

In this section:

Description

This function is one of the four elements of a simulation model. For a more detailed description on how a simulation model is constructed and calculated, please read the introduction on XLSTAT-Sim.

This tool allows to create a function that calculates a statistic after each iteration of the simulation process. The statistic is computed and stored. During the step by step simulations, you can track how the statistic evolves. In the simulation report you can optionally see details on the statistic. A wide choice of statistics is available.

To define the statistic function (physically, a cell on the Excel sheet), you need to create a call to a XLSTAT_SimStatX/TheoX/SPCX function or to use the dialog box that will generate for you the formula calling the function. X stands for the statistic as defined in the tables below. A variable based on the corresponding statistic is created.

XLSTAT_SimStat/Theo/SPC Syntax

XLSTAT_SimStatX(StatName, Reference, Visible)

XLSTAT_SimTheoX(StatName, Reference, Visible)

XLSTAT_SimSPCX(StatName, Reference, Visible)

X stands for one of the selected statistic. The available statistics are listed in the tables below.

StatName is a string that contains the name of the statistic or a reference to a cell where the name is located. The name is used during the report to identify the statistic.

Reference indicates the model variable to be tracked. This is a reference to a cell in the same Excel sheet.

Visible is an optional input that indicates if the details of this statistic should be displayed when the "Filter level for results" in the Options dialog box of XLSTAT-Sim is set to "Activated" (see the Format tab). 0 deactivates the display and 1 activates the display. Default value is 1.

Descriptive statistics

The following descriptive statistics are available:

Statistic name	XLSTAT Name
Number of observations	XLSTAT_SimStatNbrObs
Number of missing values	XLSTAT_SimStatNbrMiss
Sum of weights	XLSTAT_SimStatSumOfWeights
Minimum	XLSTAT_SimStatMinimum
Maximum	XLSTAT_SimStatMaximum
Frequency of minimum	XLSTAT_SimStatFreqMin
Frequency of maximum	XLSTAT_SimStatFreqMax
Range	XLSTAT_SimStatAmplitude
1st quartile	XLSTAT_SimStat1stQuartile
Median	XLSTAT_SimStatMedian
3rd quartile	XLSTAT_SimStat3rdQuartile
Sum	XLSTAT_SimStatSum
Mean	XLSTAT_SimStatMean
Variance n	XLSTAT_SimStatVarianceN
Variance n-1	XLSTAT_SimStatVarianceN1
Standard deviation n	XLSTAT_SimStatStdevN
Standard deviation n-1	XLSTAT_SimStatStdevN1
Variation coefficient	XLSTAT_SimStatVariation
Skewness (Pearson)	XLSTAT_SimStatSkewnessPearson
Skewness (Fisher)	XLSTAT_SimStatSkewnessFisher
Skewness (Bowley)	XLSTAT_SimStatSkewnessBowley
Kurtosis (Pearson)	XLSTAT_SimStatKurtosisPearson
Kurtosis (Fisher)	XLSTAT_SimStatKurtosisFisher
Standard error of the mean	XLSTAT_SimStatStdErrorOfMean
Lower bound on mean	XLSTAT_SimStatLowerBound95Perc
Upper bound on mean	XLSTAT_SimStatUpperBound95Perc
Standard deviation of skewness	XLSTAT_SimStatStdErrorOfSkewness
Standard deviation of kurtosis	XLSTAT_SimStatStdErrorOfKurtosis
Mean absolute deviation	XLSTAT_SimStatMeanAbsDeviation
Median absolute deviation	XLSTAT_SimStatMedianAbsDeviation
Geometric mean	XLSTAT_SimStatGeometricMean
Geometric standard deviation	XLSTAT_SimStatGSD
Harmonic mean	XLSTAT_SimStatHarmonicMean

Details and formulae relative to the above statistics are available in the description section of the "Descriptive statistics" tool of XLSTAT.

Theoretical statistics

These statistics are based on the theoretical computation of the mean, variance and standard deviation of the distribution, as opposed to the empirical computation based on the simulated samples.

Statistic name	XLSTAT Name
Mean	XLSTAT_SimTheoMean
Variance	XLSTAT_SimTheoVariance
Standard deviation	XLSTAT_SimTheoStdev

SPC

Statistics from the domain of SPC (Statistical Process Control) are listed hereunder. These statistics are only available and calculated, if you have a valid license for the XLSTAT-SPC module.

Statistic name	XLSTAT Name
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

Dialog box

OK: click this button to create the statistic.

Cancel: click this button to close the dialog box without doing any modification.

Help: click this button to display help.

 : click this button to reload the default options.

 : click this button to delete the data selections.

General tab:

Name: Enter the name of the statistic or select a cell where the name is available. If you select a cell, it depends on the selection in the options, whether an absolute (for example $A4$) or a relative reference (for example A4) to the cell is created. (See the Options section for more details).

Reference: Choose a cell in the active Excel sheet that contains the simulation model variable that you want to track with the selected statistic.

Statistic: Activate one of the following options and choose the statistic to compute:

- **Descriptive**: Select one of the available statistics (See description section for more details).
- **Theoretical**: Select one of the available statistics (See description section for more details).
- **SPC**: Select one of the available statistics (See description section for more details).
-

Display Results: Activate this option to display the detailed results for statistic in the simulation report. This option is only active, if you selected the restricted filter level in the simulation preferences. (See the XLSTAT-Sim options for more details).

Results

A function call to XLSTAT_SimStat/Theo/SPC with the selected parameters and the following syntax will be generated in the active Excel cell:

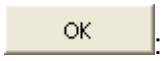
=XLSTAT_SimStat/Theo/SPC(**DistName**, **Reference**, **Visible**)

The background color and the font color in the Excel cell are applied according to your choices in the XLSTAT-Sim options.

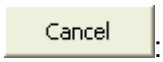
Run

Once you have designed the simulation model using the four tools "define a distribution", "define a scenario variable", "define a result", and "define a statistic", you can click the  icon of "XLSTAT-SIM" toolbar to display the "Run" dialog box that lets you define additional options before running the simulation model and displaying the report. A description of the results is also available.

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: click this button to start the calculations.



: click this button to close the dialog box without doing any calculations.

: click this button to display help.


: click this button to reload the default options.


: click this button to delete the data selections.


: click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Number of simulations: Enter the number of simulations to perform for the model (Default value: 300).

Correlation/Covariance matrix: Activate this option to include a correlation or covariance matrix in the simulation model. Column and row headers must be selected as they are used by XLSTAT to know which variables are involved. As a matter of fact, column and row labels must be identical to the names of the corresponding distribution fields of the simulation model.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if the row and column labels are selected.

Options tab:

Tornado/Spider: Choose the options for the calculation of the tornado and spider analysis.

- **Central value**: Choose how the central values around which the intervals to check during the tornado and spider analysis needs to be computed for each variable.
- **Median**: The default value of the distribution fields is the median of the simulated values.
- **Default cell value**: The default value defined for the variables is used.
- **Number of points**: Choose the number of points between the two bounds of the intervals that are used for the tornado analysis.
- **Interval definition**: Choose an option for the definition of the limits of the intervals of the variables that are checked during the tornado/spider analyses.
- **Percentile of variable**: Choose which two percentiles need to be used to determine the bounds of the intervals for the tornado/spider analyses. You can choose between [25%, 75%], [10%, 90%], and [5%, 95%]. This option is only available if the median is the central value.
- **% of deviation of value**: Choose which bounds; computed as % of the central value should be used as the bounds for the intervals. You can choose between [-25%, 25%], [-10%, 10%], and [-5%, 5%].

SPC tab:

Calculate Process capabilities: Activate this option to calculate process capabilities for input random variables, result variables and statistics.

- **Variable name s**: Select the data that correspond to the names of the variables for which you want to calculate process capabilities.
- **LSL**: Select the data that correspond to the lower specification limit (LSL) of the process for the variables for which the names have been selected.
- **USL**: Select the data that correspond to the upper specification limit (USL) of the process for the variables for which the names have been selected.
- **Target**: Select the data that correspond to the target of the process for the variables for which the names have been selected.

- **Confidence interval (%)**: If the calculation of the process capabilities is activated, please enter the percentage range of the confidence interval to use for calculating the confidence interval around the parameters. Default value: 95.

Outputs tab:

Correlations: Activate this option to display the correlation matrix between the variables. If the "**significant correlations in bold** " option is activated, the correlations that are significant at the selected significance level are displayed in bold.

- **Type of correlation**: Choose the type of correlation to use for the computations (see the description section for more details).
- **Significance level (%)**: Enter the significance level for the test of on the correlations (default value: 5%).
- **p-values**: Activate this option to display the p-values corresponding to the correlations.
- **Sensitivity**: Activate this option to display the results of the sensitivity analysis.

Tornado: Activate this option to display the results of the tornado analysis.

Spider: Activate this option to display the results of the spider analysis.

Simulation details: Activate this option to display the details on the iterations of the simulation.

Descriptive statistics: Activate this option to compute and display descriptive statistics for the variables of the model.

- **All**: Click this button to select all.
- **None**: Click this button to deselect all.
- **Display vertically**: Check this option so that the table of descriptive statistics is displayed vertically (one line per descriptive statistic).

Charts tab:

This tab is divided into three sub-tabs.

Histograms tab:

Histograms: Activate this option to display the histograms of the samples. For a theoretical distribution, the density function is displayed.

- **Bars**: Choose this option to display the histograms with a bar for each interval.
- **Continuous lines**: Choose this option to display the histograms with a continuous line.

Cumulative histograms: Activate this option to display the cumulated histograms of the samples.

Intervals: Select one of the following options to define the intervals of the histogram:

- **Number**: Choose this option to enter the number of intervals to create.
- **Width**: Choose this option to define a fixed width for the intervals.
- **User defined**: Select a column containing in increasing order the lower bound of the first interval, and the upper bound of all the intervals.
- **Minimum**: Activate this option to enter the minimum value of the histogram. If the Automatic option is chosen, the minimum is that of the sample. Otherwise, it is the value defined by the user.

Box plots tab:

Box plots: Check this option to display box plots (or box-and-whisker plots). See the description section for more details.

- **Horizontal**: Check this option to display box plots and scattergrams horizontally.
- **Vertical**: Check this option to display box plots and scattergrams vertically.
- **Group plots**: Check this option to group together the various box plots and scattergrams on the same chart to compare them.
- **Minimum/Maximum**: Check this option to systematically display the points corresponding to the minimum and maximum (box plots).
- **Outliers**: Check this option to display the points corresponding to outliers (box plots) with a hollowed-out circle.

Scattergrams: Check this option to display scattergrams. The mean (red +) and the median (red line) are always displayed.

Normal P-P plots: Check this option to display P-P plots.

Normal Q-Q Charts: Check this option to display Q-Q plots.

Correlations tab:

Correlation maps: Several visualizations of a correlation matrix are proposed.

- The "**blue-red** " option allows to represent low correlations with cold colors (blue is used for the correlations that are close to -1) and the high correlations are with hot colors (correlations close to 1 are displayed in red color).
- The "**Black and white** " option allows to either display in black the positive correlations and in white the negative correlations (the diagonal of 1s is display in grey color), or to display in black the significant correlations, and in white the correlations that are not significantly different from 0.
- The "**Patterns** " option allows to represent positive correlations by lines that rise from left to right, and the negative correlations by lines that rise from right to left. The higher the absolute value of the correlation, the large the space between the lines.

Scatter plots: Activate this option to display the scatter plots for all two by two combinations of variables.

- **Matrix of plots**: Check this option to display all possible combinations of variables in pairs in the form of a two-entry table with the various variables displayed in rows and in columns.
- **Histograms**: Activate this option so that XLSTAT displays a histogram when the X and Y variables are identical.
- **Q-Q plots**: Activate this option so that XLSTAT displays a Q-Q plot when the X and Y variables are identical.
- **Confidence ellipses**: Activate this option to display confidence ellipses. The confidence ellipses correspond to a x% confidence interval (where x is determined using the significance level entered in the General tab) for a bivariate normal distribution with the same means and the same covariance matrix as the variables represented in abscissa and ordinates.

Results

The first results are general results that display information about the model:

Distributions: This table shows for each input random variable in the model, its name, the Excel cell where it is located, the selected distribution, the static value, the data type, the truncation mode and bounds and the parameters of the distribution.

Scenario variables: This table shows for each input random variable in the model, its name, the Excel cell where it is located, the default value, the type, the lower und upper limit and the step size.

Result variables: This table shows for each result variable in the model, its name, the Excel cell where it is located, and the formula for its calculation.

Statistics: This table shows for each statistic in the model, its name, the Excel cell that contains it and the selected statistic.

Correlation/covariance matrix: If the option correlation/covariance matrix in the simulation model has been activated, then this table displays the input correlation/covariance matrix.

Convergence: If the option convergence in the simulation options has been activated, then this table displays for each result variable that has been selected for convergence checking, the value and the variation of the lower and upper bound of the confidence interval for the selected convergence criterion. Under the matrix information about the selected convergence criterion, the corresponding threshold of variation, and the number of executed iterations of simulation are displayed.

In the following section, details for the different model elements, distributions, scenario variables, result variables and statistics, are displayed.

Descriptive statistics: For each type of variable, the statistics selected in the dialog box are displayed in a table.

Descriptive statistics for the intervals: This table displays for each interval of the histogram its lower bound, upper bound, the frequency (number of values of the sample within the interval), the relative frequency (the number of values divided by the total number of values in the sample), and the density (the ratio of the frequency to the size of the interval).

Sensitivity: A table with the correlations, the contributions and the absolute value of the contributions between the input random variables is displayed for each result variable. The contributions are then plotted on a chart.

Tornado: This table displays the minimum, the maximum and the range of the result variable when the input random variables and the scenario variables vary in the defined ranges. Then the minimum and the maximum are shown on a chart.

Spider: This table displays for all the points that are evaluated during the tornado analysis the value of each result variable when the input random variables and scenario variables vary. These values are then displayed in on the spider chart.

The **correlation matrix** and the table of the p-values are displayed so that you can see the relationships between the input variables and the output variables. The correlation maps allow identifying potential structures in the matrix, of to quickly identify interesting correlations.

Simulation details: A table showing the values of each variable at each iteration is displayed.

Power analysis

Compare means (Power and sample size)

Use this tool to compute power and sample size in a statistical test comparing means. T test, z test and non parametric tests are available.

In this section:

Description

XLSTAT includes several tests to compare means, namely the t test, the z test and other non parametric tests like Mann-Whitney test. XLSTAT allows as well estimating the power of these tests and calculates the number of observations required to obtain sufficient power.

When testing a hypothesis using a statistical test, there are several decisions to take:

- The null hypothesis H0 and the alternative hypothesis Ha.
- The statistical test to use.
- The type I error also known as alpha. It occurs when one rejects the null hypothesis when it is true. It is set a priori for each test and is 5%.

The type II error or beta is less studied but is of great importance. In fact, it represents the probability that one does not reject the null hypothesis when it is false. We can not fix it upfront, but based on other parameters of the model we can try to minimize it. The power of a test is calculated as $1 - beta$ and represents the probability that we reject the null hypothesis when it is false.

We therefore wish to maximize the power of the test. XLSTAT calculates the power (and beta) when other parameters are known. For a given power, it also allows to calculate the sample size that is necessary to reach that power.

The statistical power calculations are usually done before the experiment is conducted. The main application of power calculations is to estimate the number of observations necessary to properly conduct an experiment.

XLSTAT allows to compare:

- A mean to a constant (with z-test, t-test and Wilcoxon signed rank test)
- Two means associated with paired samples (with z-test, t-test and Wilcoxon signed rank test)
- Two means associated with independent samples (with z-test, t-test and Mann-Whitney test)

We use the t-test when the variance of the population is estimated and the z-test when it is known. In each case, the parameters will be different and will be shown in the dialog box. The non parametric tests are used when the distribution assumption is not met.

Methods

The sections of this document dedicated to the t-test, the z-test and the non parametric tests describe in detail the methods themselves.

The power of a test is usually obtained by using the associated non-central distribution. Thus, for the t-test, the non-central Student distribution is used.

T-test for one sample

The power of this test is obtained using the non-central Student distribution with non-centrality parameter:

$$NCP = \left| \frac{\overline{X} - X_0}{SD} \cdot \sqrt{N} \right|$$

With X_0 the theoretical mean and SD the standard deviation.

The part $\frac{\overline{X} - X_0}{SD}$ is called the effect size.

T-test for two paired samples

The same formula as for the one sample case applies, but the standard deviation is calculated differently, we have:

$$NCP = \left| \frac{\bar{X}_1 - \bar{X}_2}{SD_{Diff}} \cdot \sqrt{N} \right|$$

with

$$SD_{Diff} = \sqrt{(SD_1^2 + SD_2^2) - 2 \cdot Corr \cdot SD_1 \cdot SD_2}$$

and $Corr$ is the correlation between the two samples.

The part $\frac{\bar{X}_1 - \bar{X}_2}{SD_{Diff}}$ is the effect size.

T-test for two independent samples

In the case of two independent samples, the standard deviation is calculated differently and we use the harmonic mean of the number of observations.

$$NCP = \left| \frac{\bar{X}_1 - \bar{X}_2}{SD_{Pooled}} \cdot \sqrt{\frac{N_{harmo}}{2}} \right|$$

with

$$SD_{Pooled} = \sqrt{\frac{(N_1 - 1) \cdot SD_1^2 - (N_2 - 1) \cdot SD_2^2}{N_1 + N_2 - 2}}$$

The part $\frac{\overline{X}_1 - \overline{X}_2}{SD_{pooled}}$ is called effect size.

Z-test for one sample

In the case of the z-test, using the classical normal distribution with a parameter added to shift the distribution.

$$NCP = \left| \frac{\overline{X} - X_0}{SD} \cdot \sqrt{N} \right|$$

With X_0 being the theoretical mean and SD being the standard deviation.

The part $\frac{\overline{X} - X_0}{SD}$ is called effect size.

Z-test for two paired samples

The same formula applies as for the one sample case, but the standard deviation is calculated differently, we have:

$$NCP = \left| \frac{\bar{X}_1 - \bar{X}_2}{SD_{Diff}} \cdot \sqrt{N} \right|$$

with

$$SD_{Diff} = \sqrt{(SD_1^2 + SD_2^2) - 2 \cdot Corr \cdot SD_1 \cdot SD_2}$$

and $Corr$ is the correlation between the two samples.

The part $\frac{\bar{X}_1-\bar{X}_2}{SD_{Diff}}$ is called effect size.

Z-test for two independent samples

In the case of two independent samples, the standard deviation is calculated differently and we use the harmonic mean of the number of observations.

$$NCP = \left| \frac{\bar{X}_1 - \bar{X}_2}{SD_{Pooled}} \cdot \sqrt{\frac{N_{harmo}}{2}} \right|$$

with

$$SD_{Pooled} = \sqrt{\frac{(N_1 - 1) \cdot SD_1^2 - (N_2 - 1) \cdot SD_2^2}{N_1 + N_2 - 2}}$$

The part $\frac{\overline{X}_1-\overline{X}_2}{SD_{pooled}}$ is called effect size.

Non parametric tests

In the case of the non parametric cases, a method called ARE (asymptotic relative efficiency) is used. This method helps to relate formulas used for the power of a t-test to those of the non parametric approaches. It has been introduced by Lehmann (1975). A factor called ARE is used. It has been shown that for mean comparisons the minimum value of the ARE is 0.864. This value is equal to 0.955 if the data are normally distributed. XLSTAT uses the minimum ARE for the computations.

To compute power of the test, the used H0 distribution is the central Student distribution: $t(N, k-2)$. The used H1 distribution is the noncentral Student distribution: $t(Nk-2, \delta)$, where the noncentrality parameter is given by:

$$\delta = d * \sqrt{(N_1 N_2 k)/(N_1 + N_2)}$$

Parameter k represents the asymptotic relative efficiency and depends on the parent distribution. Parameter d is the effect size defined like in the t-test case depending on the type of sample studied (independent, paired or one-sample).

Calculating sample size

To calculate the number of observations required, XLSTAT uses an algorithm that searches for the root of a function. It is called the Van Wijngaarden- Dekker-Brent algorithm (Brent, 1973). This algorithm is adapted to the case where the derivatives of the function are not known. It tries to find the root of:

power (N) - expected_power

We then obtain the size N such that the test has a power as close as possible to the desired power.

Effect size

This concept is very important in power calculations. Indeed, Cohen (1988) developed this concept. The effect size is a quantity that will allow calculating the power of a test without entering any parameters but will tell if the effect to be tested is weak or strong.

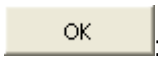
In the context of comparisons of means, the conventions of magnitude of the effect size are:

- $d = 0.2$, the effect is small.
- $d = 0.5$, the effect is moderate.
- $d = 0.8$, the effect is strong.

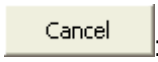
XLSTAT allows entering directly the effect size.

Dialog box

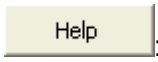
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. Below are descriptions of the various elements of the dialog box.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.



: Click this button to display help options.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Goal: Choose between computing power and sample size estimation.

test: Select the test you want to apply.

Alternative hypothesis: Select the alternative hypothesis to be tested.

Theoretical mean (when only one sample is used): Enter the value of the theoretical mean to be tested.

Alpha: Enter the value of the type I error (alpha, between 0.001 and 0.999).

Power (when sample size estimation has been selected): Enter the value of the power to be reached.

Sample size (group 1) (when power computation has been selected): Enter the size of the first sample.

Sample size (group 2) (when power computation has been selected): Enter the size of the second sample.

N1/N2 ratio (when sample size has been selected and when there are two samples): Enter the ratio between the sizes of the first and the second samples.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Effect size tab:

Effect size: Select this option to directly enter the effect size D (see the description part of this help).

Parameters: Select this option to enter the test parameters directly.

Mean (group 1): Enter the mean for group 1.

Mean (group 2): Enter the mean for group 2.

Std error (group 1): Enter the standard error for group 1.

Std error (group 2): Enter the standard error for group 2.

Correlation (when using paired samples): Enter the correlation between the groups.

Graphics tab:

Simulation plot: Activate this option if you want to plot different parameters of the test. Two parameters can vary. All remaining parameters are used as they were defined in the General tab.

X axis: Select the parameter to be used on the X axis of the simulation plot. You can either choose the power or the sample size, the type I error (alpha) or the effect size. Depending on what we are looking for, we will have on the Y axis either the power or the sample size.

Interval size: Enter the minimum, maximum and interval size for the X axis of the simulation plot.

Results

Results: This table displays the parameters of the test and the power or the required number of observations. The parameters obtained by the calculation are in bold format. An explanation is displayed below this table.

Intervals for the simulation plot: This table is composed of two columns: power and sample size or alpha depending on the parameters selected in the dialog box. It helps building the simulation plot.

Simulation plot: This plot shows the evolution of the parameters as defined in the graphics tab of the dialog box.

Example

An example of power calculation based on a test is available on XLSTAT Help Center:

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

An example of calculating the required sample size is available on the XLSTAT Help Center:

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

References

Brent R. P (1973). Algorithms for Minimization Without Derivatives. Englewood Cliffs, NJ: Prentice-Hall.

Cohen J. (19 88). 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.

Compare variances (Power and sample size)

Use this tool to compute power and sample size in a statistical test comparing variances.

In this section:

Description

XLSTAT includes several tests to compare variances. XLSTAT can also calculate the power or the number of observations required for a test based on Fisher's F distribution to compare variances.

When testing a hypothesis using a statistical test, there are several decisions to take:

- The null hypothesis H0 and the alternative hypothesis Ha.
- The statistical test to use.
- The type I error also known as alpha. It occurs when one rejects the null hypothesis when it is true. It is set a priori for each test and is 5%.

The type II error or beta is less studied but is of great importance. In fact, it represents the probability that one does not reject the null hypothesis when it is false. We can not fix it upfront, but based on other parameters of the model we can try to minimize it. The power of a test is calculated as $1 - beta$ and represents the probability that we reject the null hypothesis when it is false.

We therefore wish to maximize the power of the test. XLSTAT calculates the power (and beta) when other parameters are known. For a given power, it also allows to calculate the sample size that is necessary to reach that power.

The statistical power calculations are usually done before the experiment is conducted. The main application of power calculations is to estimate the number of observations necessary to properly conduct an experiment.

XLSTAT allows to compare two variances. The parameters are shown in the dialog box.

Methods

The sections of this document dedicated to the tests used to compare variances test describe in detail the methods themselves.

The power of a test is usually obtained by using the associated non-central distribution. In that case, we use the F distribution.

Several hypotheses can be tested, but the most common are the following (two-tailed):

- H0: The difference between the variances is equal to 0.
- Ha: The difference between the variances is different from 0.

The power computation will give the proportion of experiments that reject the null hypothesis. The calculation is done using the F distribution with the ratio of the variances as parameter and the sample sizes – 1 as degrees of freedom.

Calculating sample size

To calculate the number of observations required, XLSTAT uses an algorithm that searches for the root of a function. It is called the Van Wijngaarden- Dekker-Brent algorithm (Brent, 1973). This algorithm is adapted to the case where the derivatives of the function are not known. It tries to find the root of:

power (N) - expected_power

We then obtain the size N such that the test has a power as close as possible to the desired power.

Effect size

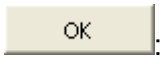
This concept is very important in power calculations. Indeed, Cohen (1988) developed this concept. The effect size is a quantity that will allow to calculate the power of a test without entering any parameters but will tell if the effect to be tested is weak or strong.

Within the comparison of variances, it is the ratio between two variances to compare.

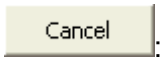
XLSTAT allows to enter directly the effect size.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. Below are descriptions of the various elements of the dialog box.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

Help: Click this button to display help options.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

General tab:

Goal: Choose between computing power and sample size estimation.

Test: Select the test you want to apply.

Alternative hypothesis: Select the alternative hypothesis to be tested.

Theoretical mean (when only one sample is used): Enter the value of the theoretical mean to be tested.

Alpha: Enter the value of the type I error (alpha, between 0.001 and 0.999).

Power (when sample size estimation has been selected): Enter the value of the power to be reached.

Sample size (group 1) (when power computation has been selected): Enter the size of the first sample.

Sample size (group 2) (when power computation has been selected): Enter the size of the second sample.

N1/N2 ratio (when sample size has been selected and when there are two samples): Enter the ratio between the sizes of the first and the second samples.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Effect size tab:

Effect size: Select this option to directly enter the effect size D (see the description part of this help).

Parameters: Select this option to enter the test parameters directly.

Variance (group 1): Enter the variance for group 1.

Variance (group 2): Enter the variance for group 2.

Graphics tab:

Simulation plot: Activate this option if you want to plot different parameters of the test. Two parameters can vary. All remaining parameters are used as they were defined in the General tab.

X axis: Select the parameter to be used on the X axis of the simulation plot. You can either choose the power or the sample size, the type I error (alpha) or the effect size. Depending on what we are looking for, we will have on the Y axis either the power or the sample size.

Interval size: Enter the minimum, maximum and interval size for the X axis of the simulation plot.

Results

Results: This table displays the parameters of the test and the power or the required number of observations. The parameters obtained by the calculation are in bold format. An explanation is displayed below this table.

Intervals for the simulation plot: This table is composed of two columns: power and sample size or alpha depending on the parameters selected in the dialog box. It helps building the simulation plot.

Simulation plot: This plot shows the evolution of the parameters as defined in the graphics tab of the dialog box.

Example

An example of power calculation based on a test is available on XLSTAT Help Center: http://www.xlstat.com/demo-pwr.htm.

An example of calculating the required sample size is available on XLSTAT Help Center: http://www.xlstat.com/demo-spl.htm.

References

Brent R. P (1973) Algorithms for Minimization Without Derivatives. Englewood Cliffs, NJ: Prentice-Hall.

Cohen J. (19 88) . Statistical Power Analysis for the Behavioral Sciences. Psychology Press, 2-nd Edition.

Compare proportions (Power and sample size)

Use this tool to compute power and sample size in a statistical test comparing proportions.

In this section:

Description

XLSTAT includes parametric tests and nonparametric tests to compare proportions. Thus we can use the z-test, chi-square test, the sign test or the McNemar test. XLSTAT can calculate the power or the number of observations necessary for these tests using either exact methods or approximations.

When testing a hypothesis using a statistical test, there are several decisions to take:

- The null hypothesis H0 and the alternative hypothesis Ha.
- The statistical test to use.
- The type I error also known as alpha. It occurs when one rejects the null hypothesis when it is true. It is set a priori for each test and is 5%.

The type II error or beta is less studied but is of great importance. In fact, it represents the probability that one does not reject the null hypothesis when it is false. We can not fix it upfront, but based on other parameters of the model we can try to minimize it. The power of a test is calculated as $1 - beta$ and represents the probability that we reject the null hypothesis when it is false.

We therefore wish to maximize the power of the test. XLSTAT calculates the power (and beta) when other parameters are known. For a given power, it also allows to calculate the sample size that is necessary to reach that power.

The statistical power calculations are usually done before the experiment is conducted. The main application of power calculations is to estimate the number of observations necessary to properly conduct an experiment.

XLSTAT allows to compare:

- A proportion to a constant (z-test with different approximations).

- Two proportions (z-test with different approximations).
- Proportions in a contingency table (chi-square test).
- Proportions in a nonparametric way (the sign test and the McNemar test)

For each case, different input parameters are used and shown in the dialog box.

Methods

The sections of this document dedicated to the tests on proportions describe in detail the methods themselves.

The power of a test is usually obtained by using the associated non-central distribution. For this specific case we will use an approximation in order to compute the power.

Comparing a proportion to a constant

The alternative hypothesis in this case is: Ha: $p_1 - p_0 \neq 0$

Various approximations are possible:

- Approximation using the normal distribution: In this case, we will use the normal distribution with means p_0 and p_1 and standard deviations

$$\sqrt{\frac{p_0(1-p_0)}{N}} \text{ and } \sqrt{\frac{p_1(1-p_1)}{N}}$$

- Exact calculation using the binomial distribution with parameters

$$\sqrt{\frac{p_0(1-p_0)}{N}} \text{ and } \sqrt{\frac{p_1(1-p_1)}{N}}$$

- Approximation using the beta distribution with parameters

$$((N-1)p_0; (N-1)(1-p_0)) \text{ and } ((N-1)p_1; (N-1)(1-p_1))$$

- Approximation using the method of the arcsin: This approximation is based on the arcsin transformation of proportions: $H(p_0)$ and $H(p_1)$. The power is obtained using the normal distribution:

$$Z_p = \sqrt{N}(H(p_0) - H(p_1)) - Z_{req}$$

with Z_{req} being the alpha-quantile of the normal distribution.

Comparing two proportions

The alternative hypothesis in this case is: Ha: $p_1 - p_2 \neq 0$

Various approximations are possible:

- Approximation using the method of the arcsin: This approximation is based on the arcsin transformation of proportions: $H(p_2)$ and $H(p_1)$. The power is obtained using the normal distribution:

$$Z_p = \sqrt{N}(H(p_2) - H(p_1)) - Z_{req}$$

with Z_{req} being the alpha-quantile of the normal distribution.

- Approximation using the normal distribution: In this case, we will use the normal distribution with means p_1 and p_2 and standard deviations:

$$\sqrt{\frac{p_1(1-p_1)}{N}} \text{ and } \sqrt{\frac{p_2(1-p_2)}{N}}$$

Chi-square test

To calculate the power of the chi-square test in the case of a contingency table 2 * 2, we use the non-central chi-square distribution with the value of the chi-square as non-centrality parameter.

It therefore seeks to see whether two groups of observations have the same behavior based on a binary variable.

We have:

	Group 1	**Group 2**
Positive	p_1	p_2
Negative	$1-p_1$	$1-p_2$

p_1, N_1 and N_2 have to be entered in the dialog box (p_2 can be found from other parameters because the test has only one degree of freedom).

Sign test

The sign test is used to see if the proportion of cases in each group is equal to 50%. It has the same principle as the one proportion test against a constant. The constant being 0.5. Power is computed using an approximation by the normal distribution or an exact method with the binomial distribution.

We must therefore enter the sample size and the proportion in one group p1 (the other proportion is such that $p_2 = 1 - p_1$).

McNemar test

The McNemar test on paired proportions is a specific case of testing a proportion against a constant. Indeed, one can represent the problem with the following table:

	Group 1	Group 2
Positive	PP	PN
Negative	NP	NN

We have PP + NN + PN + NP = 1. We want to try to see the effect of a treatment; we are therefore interested in NP and PN. The other values are not significant.

The test inputs are: Proportion P_1 = NP and Proportion P_2 = PN. With necessarily $P_1 + P_2 < 1$.

The effect is calculated only on a percentage of NP + PN of the sample. The proportion of individuals that change from negative to positive is calculated as NP / (NP + NP). So we will try to compare this figure to a value of 50% to see if we have more individuals who go from positive to negative than individuals who go from negative to positive.

This test is well suited for medical applications.

Calculating sample size

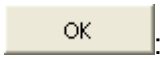
To calculate the number of observations required, XLSTAT uses an algorithm that searches for the root of a function. It is called the Van Wijngaarden- Dekker-Brent algorithm (Brent, 1973). This algorithm is adapted to the case where the derivatives of the function are not known. It tries to find the root of:

power (N) - expected_power

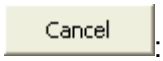
We then obtain the size N such that the test has a power as close as possible to the desired power.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. Below are descriptions of the various elements of the dialog box.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help options.


: Click this button to reload the default options.

 : Click this button to delete the data selections.

General tab:

Goal: Choose between computing power and sample size estimation.

Test: Select the test you want to apply.

Alternative hypothesis: Select the alternative hypothesis to be tested.

Theoretical mean (when only one sample is used): Enter the value of the theoretical mean to be tested.

Alpha: Enter the value of the type I error (alpha, between 0.001 and 0.999).

Power (when sample size estimation has been selected): Enter the value of the power to be reached.

Sample size (group 1) (when power computation has been selected): Enter the size of the first sample.

Sample size (group 2) (when power computation has been selected): Enter the size of the second sample.

N1/N2 ratio (when sample size has been selected and when there are two samples): Enter the ratio between the sizes of the first and the second samples.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Effect size tab:

Proportion 1: Enter the proportion for group 1.

Proportion 2: Enter the proportion for group 2.

Graphics tab:

Simulation plot: Activate this option if you want to plot different parameters of the test. Two parameters can vary. All remaining parameters are used as they were defined in the General tab.

X axis: Select the parameter to be used on the X axis of the simulation plot. You can either choose the power or the sample size, the type I error (alpha) or the effect size. Depending on what we are looking for, we will have on the Y axis either the power or the sample size.

Interval size: Enter the minimum, maximum and interval size for the X axis of the simulation plot.

Results

Results: This table displays the parameters of the test and the power or the required number of observations. The parameters obtained by the calculation are in bold format. An explanation is displayed below this table.

Intervals for the simulation plot: This table is composed of two columns: power and sample size or alpha depending on the parameters selected in the dialog box. It helps building the simulation plot.

Simulation plot: This plot shows the evolution of the parameters as defined in the graphics tab of the dialog box.

Example

An example of power calculation based on a test is available on XLSTAT Help Center: http://www.xlstat.com/demo-pwr.htm.

An example of calculating the required sample size is available on XLSTAT Help Center: http://www.xlstat.com/demo-spl.htm

References

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.

Compare correlations (Power and sample size)

Use this tool to compute power and sample size in a statistical test comparing Pearson correlations.

In this section:

Description

XLSTAT offers a test to compare correlations. XLSTAT can calculate the power or the number of observations necessary for this test.

When testing a hypothesis using a statistical test, there are several decisions to take:

- The null hypothesis H0 and the alternative hypothesis Ha.
- The statistical test to use.
- The type I error also known as alpha. It occurs when one rejects the null hypothesis when it is true. It is set a priori for each test and is 5%.

The type II error or beta is less studied but is of great importance. In fact, it represents the probability that one does not reject the null hypothesis when it is false. We can not fix it upfront, but based on other parameters of the model we can try to minimize it. The power of a test is calculated as $1 - beta$ and represents the probability that we reject the null hypothesis when it is false.

We therefore wish to maximize the power of the test. XLSTAT calculates the power (and beta) when other parameters are known. For a given power, it also allows to calculate the sample size that is necessary to reach that power.

The statistical power calculations are usually done before the experiment is conducted. The main application of power calculations is to estimate the number of observations necessary to properly conduct an experiment.

XLSTAT allows to compare:

- One correlation to 0.
- One correlation to a constant.

- Two correlations.

Methods

The section of this document dedicated to the correlation tests describes in detail the methods themselves.

The power of a test is usually obtained by using the associated non-central distribution. For this specific case we will use an approximation in order to compute the power.

Comparing on correlation to 0

The alternative hypothesis in this case is: Ha: $r \neq 0$

The method used is an exact method based on the non-central Student distribution.

The non-centrality parameter used is the following:

$$NCP = \sqrt{\frac{r^2}{1-r^2}} \cdot \sqrt{N}$$

The part $\frac{r^2}{1-r^2}$ is called effect size.

Comparing one correlation to a constant

The alternative hypothesis in this case is: Ha: $r \neq r_0$

The power calculation is done using an approximation by the normal distribution. We use the Fisher Z-transformation:

$$Z_r = \frac{1}{2}\log(\frac{1+r}{1-r})$$

The effect size is: $Q = |Z_r - Z_{r_0}|$

The power is then found using the area under the curve of the normal distribution to the left of Zp:

$Z_p = Q \cdot \sqrt{N-3} - Z_{req}$ where Z_{req} is the quantile of the normal distribution for alpha.

Comparing two correlations

The alternative hypothesis in this case is: Ha: $r_1 - r_2 \neq 0$

The power calculation is done using an approximation by the normal distribution. We use the Fisher Z-transformation:

$$Z_r = \frac{1}{2}\log(\frac{1+r}{1-r})$$

The effect size is: $Q = |Z_{r_1} - Z_{r_2}|$

The power is then found using the area under the curve of the normal distribution to the left of Z_p:

$$Z_p = Q \cdot \sqrt{\frac{N' - 3}{2}} - Z_{req}$$

where Z_{req} is the quantile of the normal distribution for alpha and

$$N' = \frac{2(N_1 - 3)(N_2 - 3)}{N_1 + N_2 - 6} + 3$$

Calculating sample size

To calculate the number of observations required, XLSTAT uses an algorithm that searches for the root of a function. It is called the Van Wijngaarden- Dekker-Brent algorithm (Brent, 1973). This algorithm is adapted to the case where the derivatives of the function are not known. It tries to find the root of:

power (N) - expected_power

We then obtain the size N such that the test has a power as close as possible to the desired power.

Effect size

This concept is very important in power calculations. Indeed, Cohen (1988) developed this concept. The effect size is a quantity that will allow to calculate the power of a test without entering any parameters but will tell if the effect to be tested is weak or strong.

In the context of comparisons of correlations conventions of magnitude of the effect size are:

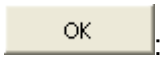
- $Q = 0.1$, the effect is small.
- $Q = 0.3$, the effect is moderate.
- $Q = 0.5$, the effect is strong.

XLSTAT allows to enter directly the effect size

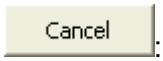
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. Below are descriptions of the various elements of

the dialog box.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help options.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Goal: Choose between computing power and sample size estimation.

Test: Select the test you want to apply.

Alternative hypothesis: Select the alternative hypothesis to be tested.

Theoretical mean (when only one sample is used): Enter the value of the theoretical mean to be tested.

Alpha: Enter the value of the type I error (alpha, between 0.001 and 0.999).

Power (when sample size estimation has been selected): Enter the value of the power to be reached.

Sample size (group 1) (when power computation has been selected): Enter the size of the first sample.

Sample size (group 2) (when power computation has been selected): Enter the size of the second sample.

N1/N2 ratio (when sample size has been selected and when there are two samples): Enter the ratio between the sizes of the first and the second samples.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Effect size tab:

Effect size: Select this option to directly enter the effect size D (see the description part of this help).

Parameters: Select this option to enter the test parameters directly.

Correlation (group 1): Enter the correlation for group 1.

Correlation (group 2): Enter the correlation for group 2.

Graphics tab:

Simulation plot: Activate this option if you want to plot different parameters of the test. Two parameters can vary. All remaining parameters are used as they were defined in the General tab.

X axis: Select the parameter to be used on the X axis of the simulation plot. You can either choose the power or the sample size, the type I error (alpha) or the effect size. Depending on what we are looking for, we will have on the Y axis either the power or the sample size.

Interval size: Enter the minimum, maximum and interval size for the X axis of the simulation plot.

Results

Results: This table displays the parameters of the test and the power or the required number of observations. The parameters obtained by the calculation are in bold format. An explanation is displayed below this table.

Intervals for the simulation plot: This table is composed of two columns: power and sample size or alpha depending on the parameters selected in the dialog box. It helps building the simulation plot.

Simulation plot: This plot shows the evolution of the parameters as defined in the graphics tab of the dialog box.

Example

An example of power calculation based on a test is available on XLSTAT Help Center: http://www.xlstat.com/demo-pwr.htm.

An example of calculating the required sample size is available on XLSTAT Help Center: http://www.xlstat.com/demo-spl.htm.

References

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.

Linear regression (Power and sample size)

Use this tool to compute power and necessary sample size in linear regression model.

In this section:

Description

XLSTAT offers a tool to apply a linear regression model. XLSTAT estimates the power or calculates the necessary number of observations associated with variations of R^2 in the framework of a linear regression.

When testing a hypothesis using a statistical test, there are several decisions to take:

- The null hypothesis H0 and the alternative hypothesis Ha.
- The statistical test to use.
- The type I error also known as alpha. It occurs when one rejects the null hypothesis when it is true. It is set a priori for each test and is 5%.

The type II error or beta is less studied but is of great importance. In fact, it represents the probability that one does not reject the null hypothesis when it is false. We can not fix it upfront, but based on other parameters of the model we can try to minimize it. The power of a test is calculated as $1 - beta$ and represents the probability that we reject the null hypothesis when it is false.

We therefore wish to maximize the power of the test. XLSTAT calculates the power (and beta) when other parameters are known. For a given power, it also allows to calculate the sample size that is necessary to reach that power.

The statistical power calculations are usually done before the experiment is conducted. The main application of power calculations is to estimate the number of observations necessary to properly conduct an experiment.

XLSTAT allows to compare:

- R^2 value to 0.
- Increase in R^2 value when new predictors are added to the model to 0.

It means testing the following hypothesis:

- H0: $R^2 = 0$ / Ha: $R^2 \neq 0$
- H0: Increase in R^2 is equal to 0 / Ha: Increase in R^2 is different from 0.

Effect size

This concept is very important in power calculations. Indeed, Cohen (1988) developed this concept. The effect size is a quantity that will allow to calculate the power of a test without entering any parameters but will tell if the effect to be tested is weak or strong.

In the context of a linear regression, conventions of magnitude of the effect size f^2 are:

- $f^2 = 0.02$, the effect is small.
- $f^2 = 0.15$, the effect is moderate.
- $f^2 = 0.35$, the effect is strong.

XLSTAT allows to enter directly the effect size but also allows to enter parameters of the model that will help calculating the effect size. We detail the calculations below:

- Using variances: We can use the variances of the model to define the size of the effect. With var_{exp} being the variance explained by the explanatory variables that we wish to test and var_{error} being the variance of the error or residual variance, we have:

$$f^2 = \frac{var_{exp}}{var_{error}}$$

- Using the R^2 (in the case H0: $R^2 = 0$): We enter the estimated square multiple correlation value (ρ^2) to define the size of the effect. We have:

$$f^2 = \frac{\rho^2}{1 - \rho^2}$$

- Using the partial R^2 (in the case H0: Increase in $R^2 = 0$): We enter the partial R^2 that is the expected difference in R^2 when adding predictors to the model to define the size of the effect. We have:

$$f^2 = \frac{R_{part}2}{1 - R_{part}2}$$

- Using the correlations between predictors (in the case H0: $R^2 = 0$): One must then select a vector containing the correlations between the explanatory variables and the dependent variable $Corr_Y$, and a square matrix containing the correlations between the explanatory variables $Corr_X$. We have:

$$f^2 = \frac{Corr_Y^t(Corr_X^{-1})Corr_Y}{1 - Corr_Y^t(Corr_X^{-1})Corr_Y}$$

Once the effect size is defined, power and necessary sample size can be computed.

Methods

The section of this document dedicated to the linear regression describes in detail the method.

The power of a test is usually obtained by using the associated non-central distribution. For this specific case we will use the Fisher non-central distribution to compute the power.

The power of this test is obtained using the non-central Fisher distribution with degrees of freedom equal to: DF1 is the number of tested variables; DF2 is the sample size from which the total number of explanatory variables included in model plus one is subtracted and the non-centrality parameter is:

$$NCP = f^2 N$$

Calculating sample size

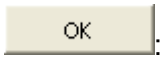
To calculate the number of observations required, XLSTAT uses an algorithm that searches for the root of a function. It is called the Van Wijngaarden- Dekker-Brent algorithm (Brent, 1973). This algorithm is adapted to the case where the derivatives of the function are not known. It tries to find the root of:

power (N) - expected_power

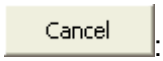
We then obtain the size N such that the test has a power as close as possible to the desired power.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. Below are descriptions of the various elements of the dialog box.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help options.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Goal: Choose between computing power and sample size estimation.

Test: Select the test you want to apply.

Alpha: Enter the value of the type I error (alpha, between 0.001 and 0.999).

Power (when sample size estimation has been selected): Enter the value of the power to be reached.

Sample size (when power computation has been selected): Enter the size of the first sample.

Number of tested predictors: Enter the number of predictors to be tested.

Total number of predictors (when testing H0: Increase in R^2=0): Enter the total number of predictors included in the model.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Effect size tab:

Determine effect size: Select the way effect size is computed.

Effect size f^2 (when effect size is entered directly): Enter the effect size (see the description part of the help for more details).

Explained variance (when effect size is computed from variances): Enter the explained variance by the tested predictors.

Error variance (when effect size is computed from variances): Enter the residual variance of the global model.

Partial R^2 (when effect size is computed using the direct approach): Enter the expected increase in R^2 when new covariates are added to the model.

rho² (when effect size is computed using the R^2): Enter the expected theoretical value of the R^2 .

The next fields appears when the hypothesis to be tested is H0: R^2=0 and when effect size is computes with the correlations between predictors.

Correlations with Ys: Select a column corresponding to the correlations between the predictors and the response variable Y. This vector must have a number of lines equal to the number of explanatory variables. Do not select the text of the column but only the numerical values.

Correlations between predictors: Select a Table corresponding to the correlations between the explanatory variables. This table should be symmetrical, have 1 on the diagonal and have a number of rows and columns equal to the number of explanatory variables. Do not select the labels of the columns or of the rows, but only the numerical values.

Graphics tab:

Simulation plot: Activate this option if you want to plot different parameters of the test. Two parameters can vary. All remaining parameters are used as they were defined in the General tab.

X axis: Select the parameter to be used on the X axis of the simulation plot. You can either choose the power or the sample size, the type I error (alpha) or the effect size. Depending on what we are looking for, we will have on the Y axis either the power or the sample size.

Interval size: Enter the minimum, maximum and interval size for the X axis of the simulation plot.

Results

Inputs: This table displays the parameters used to compute effect size.

Results: This table displays the alpha, the effect size and the power or the required number of observations. The parameters obtained by the calculation are in bold format. An explanation is displayed below this table.

Intervals for the simulation plot: This table is composed of two columns: power and sample size or alpha depending on the parameters selected in the dialog box. It helps building the simulation plot.

Simulation plot: This plot shows the evolution of the parameters as defined in the graphics tab of the dialog box.

Example

An example of power calculation based on a test is available on XLSTAT Help Center: http://www.xlstat.com/demo-pwr.htm.

An example of calculating the required sample size is available on XLSTAT Help Center: http://www.xlstat.com/demo-spl.htm.

References

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 (Power and sample size)

Use this tool to compute power and necessary sample size in analysis of variance, repeated measures analysis of variance or analysis of covariance model.

In this section:

Description

XLSTAT offers tools to apply analysis of variance (ANOVA), repeated measures analysis of variance and analysis of covariance (ANCOVA). XLSTAT estimates the power or calculates the necessary number of observations associated with these models.

When testing a hypothesis using a statistical test, there are several decisions to take:

- The null hypothesis H0 and the alternative hypothesis Ha.
- The statistical test to use.
- The type I error also known as alpha. It occurs when one rejects the null hypothesis when it is true. It is set a priori for each test and is 5%.

The type II error or beta is less studied but is of great importance. In fact, it represents the probability that one does not reject the null hypothesis when it is false. We can not fix it upfront, but based on other parameters of the model we can try to minimize it. The power of a test is calculated as $1 - beta$ and represents the probability that we reject the null hypothesis when it is false.

We therefore wish to maximize the power of the test. XLSTAT calculates the power (and beta) when other parameters are known. For a given power, it also allows to calculate the sample size that is necessary to reach that power.

The statistical power calculations are usually done before the experiment is conducted. The main application of power calculations is to estimate the number of observations necessary to properly conduct an experiment.

XLSTAT can therefore test:

- In the case of a one-way ANOVA or more fixed factors and interactions, as well as in the case of ANCOVA:

- H0: The means of the groups of the tested factor are equal.
- Ha: At least one of the means is different from another.
- In the case of repeated measures ANOVA for a within-subjects factor:
- H0: The means of the groups of the within subjects factor are equal.
- Ha: At least one of the means is different from another.
- In the case of repeated measures ANOVA for a between-subjects factor:
- H0: Les The means of the groups of the between subjects factor are equal.
- Ha: At least one of the means is different from another.
- In the case of repeated measures ANOVA for an interaction between a within-subjects factor and a between-subjects factor:
- H0: The means of the groups of the within-between subjects interaction are equal.
- Ha: At least one of the means is different from another:

Effect size

This concept is very important in power calculations. Indeed, Cohen (1988) developed this concept. The effect size is a quantity that will allow to calculate the power of a test without entering any parameters but will tell if the effect to be tested is weak or strong.

In the context of an ANOVA-type model, conventions of magnitude of the effect size f are:

- $f = 0.1$, the effect is small.
- $f = 0.25$, the effect is moderate.
- $f = 0.4$, the effect is strong.

XLSTAT allows to directly enter the effect size but also allows you to enter parameters of the model that will calculate the effect size. We detail the calculations below:

- Using variances: We can use the variances of the model to define the size of the effect. With var_{exp} being the variance explained by the explanatory factors that we wish to test and var_{error} being the variance of the error or residual variance, we have:

$$f^2 = \frac{var_{exp}}{var_{error}}$$

- Using the direct approach: We enter the estimated value of η^2 which is the ratio between the explained variance by the studied factor and the total variance of the model. For more details on η^2, please refer to Cohen (1988, chap. 8.2). We have:

$$f = \sqrt{\frac{\eta^2}{1-\eta^2}}$$

- Using the means of each group (in the case of one-way ANOVA or within subjects repeated measures ANOVA): We select a vector with the averages for each group. It is also possible to have groups of different sizes, in this case, you must also select a vector with different sizes (the standard option assumes that all groups have equal size). We have:

$$f = \frac{\sqrt{\sum_i \frac{(mi-m)^2}{k}}}{SD_{intra}}$$

with m_i mean of group i, m mean of the means, k number of groups and SD_{intra} within-group standard deviation.

- When an ANCOVA is performed, a term has to be added to the model in order to take into account the quantitative predictors. The effect size is then multiplied by

$$\sqrt{\frac{1}{1-\rho^2}}$$

where ρ^2 is the theoretical value of the square multiple correlation coefficient associated to the quantitative predictors.

Once the effect size is defined, power and necessary sample size can be computed.

Methods

The section of this document dedicated to the different methods describes in detail the methods themselves.

The power of a test is usually obtained by using the associated non-central distribution. For this specific case we will use the Fisher non-central distribution to compute the power.

We first introduce some notations:

- NbrGroup: Number of groups we wish to test.
- N: Sample size.
- NumeratorDF: Numerator degrees of freedom for the F distribution (see bellow for more details).
- NbrRep: Number of repetition (measures) for repeated measures ANOVA.
- ρ: Correlation between measures for repeated measures ANOVA.
- ϵ: Geisser-Greenhouse non sphericity correction.
- NbrPred: Number of predictors in an ANCOVA model.

For each method, we give the first and second degrees of freedom and the non-centrality parameter:

- One-way ANOVA:

$$DF1 = \text{NbrGroup} - 1 \quad DF2 = N - NbrGroup \quad NCP = f^2 N$$

- ANOVA with fixed effects and interactions:

$$DF1 = NumeratorDF \quad DF2 = N - NbrGroup \quad NCP = f^2 N$$

- Repeated measures ANOVA within-subjects factor:

$$DF1 = NbrRep - 1 \quad DF2 = (N - NbrGroup)(NbrRep - 1) \quad NCP = f^2 \frac{N.NbrRep}{1-\rho}\epsilon$$

- Repeated measures ANOVA between-subjects factor:

$$DF1 = NbrGroup - 1 \quad DF2 = N - NbrGroup \quad NCP = f^2 \frac{N.NbrRep}{1-\rho(NbrRep-1)}\epsilon$$

- Repeated measures ANOVA interaction between a within-subject factor and a between-subject factor:

$$DF1 = (NbrRep - 1)(NbrGroup - 1) \quad DF2 = (N - NbrGroup)(NbrRep - 1) \quad NCP =$$

- ANCOVA:

$$DF1 = NumeratorDF \quad DF2 = N - NbrGroup - NbrPredictors \quad NCP = f^2 N$$

Calculating sample size

To calculate the number of observations required, XLSTAT uses an algorithm that searches for the root of a function. It is called the Van Wijngaarden- Dekker-Brent algorithm (Brent, 1973). This algorithm is adapted to the case where the derivatives of the function are not known. It tries to find the root of:

power (N) - expected_power

We then obtain the size N such that the test has a power as close as possible to the desired power.

Numerator degrees of freedom

In the framework of an ANOVA with fixed factor and interactions or an ANCOVA, XLSTAT proposes to enter the number of degrees of freedom for the numerator of the non-central F distribution. This is due to the fact that many different models can be tested and computing numerator degrees of freedom is a simple way to test all kind of models.

Practically, the numerator degrees of freedom is equal to the number of group associated to the factor minus one in the case of a fixed factor. When interactions are studied, it is equal to the product of the degrees of freedom associated to each factor included in the interaction.

Suppose we have a 3-factor model, A (2 groups), B (3 groups), C (3 groups), 3 second order interactions AB, AC and BC and one third-order interaction ABC We have 18 groups.

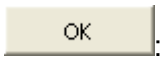
To test the main effects A, we have: $NbGroups = 18$ and N$umeratorDF = (2-1) = 1$.

To test the interactions, eg AB, we have $NbGroups = 18$ and $NumeratorDF = (2-1)(3-1) = 2$. If you wish to test the third order interaction (ABC), we have $NbGroups = 18$ and $NumeratorDF = (2-1)(3-1)(3-1) = 4$.

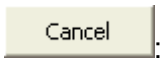
In the case of an ANCOVA, the calculations will be similar.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. Below are descriptions of the various elements of the dialog box.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help options.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Goal: Choose between computing power and sample size estimation.

Test: Select the test you want to apply.

Alpha: Enter the value of the type I error (alpha, between 0.001 and 0.999).

Power (when sample size estimation has been selected): Enter the value of the power to be reached.

Sample size (when power computation has been selected): Enter the size of the first sample.

Number of groups: Enter the total number of groups included in the model.

Number of tested predictors: Enter the number of predictors to be tested.

NumDF: Enter the number of degrees of freedom associated to the tested factor (Number of groups -1 in the case of a first order factor). For more details, see the description part of this help.

Correlation between measures: Enter the correlation between measures for repeated measures ANOVA.

Sphericity correction: Enter the Geisser-Greenhouse epsilon for correction of non-sphericity for repeated measures ANOVA. If the hypothesis of sphericity is not rejected, then epsilon=1.

Number of tested predictors: Enter the number of predictors in the ANCOVA model.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Effect size tab:

Determine effect size: Select the way effect size is computed.

Effect size f (when effect size is entered directly): Enter the effect size (see the description part of the help for more details).

Explained variance (when effect size is computed from variances): Enter the explained variance by the tested factors.

Error variance (when effect size is computed from variances): Enter the residual variance of the global model.

Within-group variance (when effect size is computed from variances): Enter the within-group variance of the model.

Partial eta² (when effect size is computed using the direct approach): Enter the expected value of η^2. For more details, see the description part of this help.

Within-group standard deviation (when effect size is computed using the means): Enter the expected within-group standard deviation of the model.

The next fields appears when applying a one-way ANOVA or repeated measures ANOVA for a between-subject factor.

Means: Select a column corresponding to the means of the groups. This vector must have a number of lines equal to the number of measures (or repetition). Do not select the label of the column but only the numerical values.

Unequal group size: Activate this option if the groups have unequal sizes. When activated, select a vector corresponding to the group sizes. This vector must have a number of lines equal to the number of measures (or repetition). Do not select the label of the column but only the numerical values. This option cannot be reached when required sample size is estimated.

Graphics tab:

Simulation plot: Activate this option if you want to plot different parameters of the test. Two parameters can vary. All remaining parameters are used as they were defined in the General tab.

X axis: Select the parameter to be used on the X axis of the simulation plot. You can either choose the power or the sample size, the type I error (alpha) or the effect size. Depending on what we are looking for, we will have on the Y axis either the power or the sample size.

Interval size: Enter the minimum, maximum and interval size for the X axis of the simulation plot.

Results

Inputs: This table displays the parameters used to compute effect size.

Results: This table displays the alpha, the effect size and the power or the required number of observations. The parameters obtained by the calculation are in bold format. An explanation is displayed below this table.

Intervals for the simulation plot: This table is composed of two columns: power and sample size or alpha depending on the parameters selected in the dialog box. It helps building the simulation plot.

Simulation plot: This plot shows the evolution of the parameters as defined in the graphics tab of the dialog box.

Example

An example of power calculation based on a test is available on XLSTAT Help Center: http://www.xlstat.com/demo-pwr.htm.

An example of calculating the required sample size is available on the XLSTAT Help Center: http://www.xlstat.com/demo-spl.htm.

References

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.

Logistic regression (Power and sample size)

Use this tool to compute power and necessary sample size in a logistic regression model.

In this section:

Description

XLSTAT offers a tool to apply logistic regression. XLSTAT estimates the power or calculates the necessary number of observations associated with this model.

When testing a hypothesis using a statistical test, there are several decisions to take:

- The null hypothesis H0 and the alternative hypothesis Ha.
- The statistical test to use.
- The type I error also known as alpha. It occurs when one rejects the null hypothesis when it is true. It is set a priori for each test and is 5%.

The type II error or beta is less studied but is of great importance. In fact, it represents the probability that one does not reject the null hypothesis when it is false. We can not fix it upfront, but based on other parameters of the model we can try to minimize it. The power of a test is calculated as $1 - beta$ and represents the probability that we reject the null hypothesis when it is false.

We therefore wish to maximize the power of the test. XLSTAT calculates the power (and beta) when other parameters are known. For a given power, it also allows to calculate the sample size that is necessary to reach that power.

The statistical power calculations are usually done before the experiment is conducted. The main application of power calculations is to estimate the number of observations necessary to properly conduct an experiment.

In the general framework of logistic regression model, the goal is to explain and predict the probability P that an event appends (usually Y=1). P is equal to:

$$P = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}$$

We have:

$$\log(\frac{P}{1-P}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

The test used in XLSTAT is based on the null hypothesis that the β_1 coefficient is equal to 0. That means that the X_1 explanatory variable has no effect on the model. For more details on logistic regression, please see the associated chapter of this help.

The hypothesis to be tested is:

- H0 : $\beta_1 = 0$
- Ha : $\beta_1 \neq 0$

Power is computed using an approximation which depends on the type of variable.

If X_1 is quantitative and has a normal distribution, the parameters of the approximation are:

- P0 (baseline probability): The probability that $Y = 1$ when all explanatory variables are set to their mean value.
- P1(alternative probability): The probability that X_1 be equal to one standard error above its mean value, all other explanatory variables being at their mean value.
- Odds ratio: The ratio between the probability that $Y = 1$, when X_1 is equal to one standard deviation above its mean and the probability that $Y = 1$ when X_1 is at its mean value.
- The R^2 obtained with a regression between X_1 and all the other explanatory variables included in the model.

If X_1 is binary and follow a binomial distribution. Parameters of the approximation are:

- P0 (baseline probability): The probability that $Y = 1$ when $X_1 = 0$.
- P1(alternative probability): The probability that $Y = 1$ when $X_1 = 1$.
- Odds ratio: The ratio between the probability that $Y = 1$, when $X_1 = 1$ and the probability that $Y = 1$ when $X_1 = 0$.
- The R^2 obtained with a regression between X_1 and all the other explanatory variables included in the model.
- The percentage of observations with $X_1 = 1$.

These approximations depend on the normal distribution.

Calculating sample size

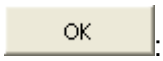
To calculate the number of observations required, XLSTAT uses an algorithm that searches for the root of a function. It is called the Van Wijngaarden- Dekker-Brent algorithm (Brent, 1973). This algorithm is adapted to the case where the derivatives of the function are not known. It tries to find the root of:

power (N) - expected_power

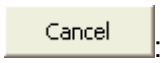
We then obtain the size N such that the test has a power as close as possible to the desired power.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. Below are descriptions of the various elements of the dialog box.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help options.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Goal: Choose between computing power and sample size estimation.

Alpha: Enter the value of the type I error (alpha, between 0.001 and 0.999).

Power (when sample size estimation has been selected): Enter the value of the power to be reached.

Sample size (when power computation has been selected): Enter the size of the first sample.

Baseline probability (P0): Enter the probability that $Y = 1$ when all explanatory variables are at their mean value or are equal to 0 when binary.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Effect size tab:

Determine effect size: Select the way effect size is computed.

Alternative probability (P1): Enter the probability that $Y = 1$ when X_1 is equal to one standard deviation above its mean value or is equal to 0 when binary.

Odds ratio: Enter the odds ratio (see the description part of this help).

R^2 of X_1 with other Xs: Enter the R^2 obtained with a regression between X_1 and the other explanatory variables of the model.

Type of variable: Select the type of variable X_1 to be analyzed (quantitative with normal distribution or binary).

Percent of N with X_1=1: In the case of a binary X_1, enter the percentage of observations with $X_1 = 1$.

Graphics tab:

Simulation plot: Activate this option if you want to plot different parameters of the test. Two parameters can vary. All remaining parameters are used as they were defined in the General tab.

X axis: Select the parameter to be used on the X axis of the simulation plot. You can either choose the power or the sample size, the type I error (alpha) or the effect size. Depending on what we are looking for, we will have on the Y axis either the power or the sample size.

Interval size: Enter the minimum, maximum and interval size for the X axis of the simulation plot.

Results

Inputs: This table displays the parameters used to compute power and required sample size.

Results: This table displays the alpha and the power or the required number of observations. The parameters obtained by the calculation are in bold format. An explanation is displayed below this table.

Intervals for the simulation plot: This table is composed of two columns: power and sample size or alpha depending on the parameters selected in the dialog box. It helps building the simulation plot.

Simulation plot: This plot shows the evolution of the parameters as defined in the graphics tab of the dialog box.

Example

An example of power calculation based on a test is available on XLSTAT Help Center: http://www.xlstat.com/demo-pwr.htm.

An example of calculating the required sample size is available on XLSTAT Help Center: http://www.xlstat.com/demo-spl.htm.

References

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.

Cox model (Power and sample size)

Use this tool to compute power and necessary sample size in a Cox proportional hazards ratio model to treat failure time data with covariates.

In this section:

Description

XLSTAT offers a tool to apply the proportional hazards ratio Cox regression model. XLSTAT estimates the power or calculates the necessary number of observations associated with this model.

When testing a hypothesis using a statistical test, there are several decisions to take:

- The null hypothesis H0 and the alternative hypothesis Ha.
- The statistical test to use.
- The type I error also known as alpha. It occurs when one rejects the null hypothesis when it is true. It is set a priori for each test and is 5%.

The type II error or beta is less studied but is of great importance. In fact, it represents the probability that one does not reject the null hypothesis when it is false. We can not fix it upfront, but based on other parameters of the model we can try to minimize it. The power of a test is calculated as $1 - beta$ and represents the probability that we reject the null hypothesis when it is false.

We therefore wish to maximize the power of the test. XLSTAT calculates the power (and beta) when other parameters are known. For a given power, it also allows to calculate the sample size that is necessary to reach that power.

The statistical power calculations are usually done before the experiment is conducted. The main application of power calculations is to estimate the number of observations necessary to properly conduct an experiment.

The Cox model is based on the hazard function which is the probability that an individual will experience an event (for example, death) within a small time interval, given that the individual has survived up to the beginning of the interval. It can therefore be interpreted as the risk of

dying at time t. The hazard function (denoted by $\lambda(t, X)$ can be estimated using the following equation:

$$\lambda(t, X) = \lambda_0(t)exp(\beta_1 X_1 + \ldots + \beta_p X_p))$$

The first term depends only on time and the second one depends on X. We are only interested by the second term. If all β_i are equal to zero then there is no hazard factor. The goal of the Cox model is to focus on the relations between the β_i and the hazard function.

The test used in XLSTAT is based on the null hypothesis that the β_1 coefficient is equal to 0. That means that the X_1 covariate is not a hazard factor. For more details on Cox model, please see the associated chapter of this help.

The hypothesis to be tested is:

- H0 : $\beta_1 = 0$
- Ha : $\beta_1 \neq 0$

A Wald statistic is computed for the test:

$$z = \frac{\beta_1}{\sqrt{\text{var}(\beta_1)}}$$

Power is computed using an approximation that depends on the normal distribution. Other parameters used in this approximation are: the event rate, which is the proportion of uncensored individuals, the standard deviation of X_1, the expected value of β_1 known as B(log(hazard ratio)) and the R^2 obtained with the regression between X_1 and the other covariates included in the Cox model.

Calculating sample size

To calculate the number of observations required, XLSTAT uses an algorithm that searches for the root of a function. It is called the Van Wijngaarden- Dekker-Brent algorithm (Brent, 1973). This algorithm is adapted to the case where the derivatives of the function are not known. It tries to find the root of:

power (N) - expected_power

We then obtain the size N such that the test has a power as close as possible to the desired power.

Calculating B

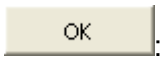
The B(log(hazard ratio)) is an estimation of the coefficient β_1 of the following equation:

$$\log(\frac{\lambda(t|X)}{\lambda_0(t)}) = \beta_1 X_1 + \ldots + \beta_k X_k$$

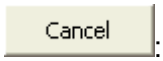
β_1 is the change in logarithm of the hazard ratio when X_1 is incremented of one unit (all other explanatory variables remaining constant). We can use the hazard ratio instead of the log. For a hazard ratio of 2, we will have $B = ln(2) = 0.693$.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. Below are descriptions of the various elements of the dialog box.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help options.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Goal: Choose between computing power and sample size estimation.

Alpha: Enter the value of the type I error (alpha, between 0.001 and 0.999).

Power (when sample size estimation has been selected): Enter the value of the power to be reached.

Sample size (when power computation has been selected): Enter the size of the first sample.

Event rate: Enter the event rate (uncensored units rate).

B(log(Hazard ratio): Enter the estimation of the parameter B associated to X1 in the Cox model.

Standard error of X1: Enter the standard error of X1.

R^2 **of X1 with other Xs**: Enter the R^2 obtained with a regression between X_1 and the other explanatory variables of the model.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Graphics tab:

Simulation plot: Activate this option if you want to plot different parameters of the test. Two parameters can vary. All remaining parameters are used as they were defined in the General tab.

X axis: Select the parameter to be used on the X axis of the simulation plot. You can either choose the power or the sample size, the type I error (alpha) or the effect size. Depending on what we are looking for, we will have on the Y axis either the power or the sample size.

Interval size: Enter the minimum, maximum and interval size for the X axis of the simulation plot.

Results

Inputs: This table displays the parameters used to compute power and required sample size.

Results: This table displays the alpha and the power or the required number of observations. The parameters obtained by the calculation are in bold format. An explanation is displayed below this table.

Intervals for the simulation plot: This table is composed of two columns: power and sample size or alpha depending on the parameters selected in the dialog box. It helps building the simulation plot.

Simulation plot: This plot shows the evolution of the parameters as defined in the graphics tab of the dialog box.

Example

An example of power calculation based on a test is available on XLSTAT Help Center: http://www.xlstat.com/demo-pwr.htm.

An example of calculating the required sample size is available on XLSTAT Help Center: http://www.xlstat.com/demo-spl.htm.

References

Brent R. P (1973) Algorithms for Minimization Without Derivatives. Englewood Cliffs, NJ: Prentice-Hall.

Cohen J. (19 88) . 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.

Sample size for clinical trials (Power and sample size)

Use this tool to compute sample size and power for different kind of clinical trials: equivalence trial, non-inferiority trial and superiority trial.

In this section:

Description

XLSTAT enables you to compute the necessary sample size for a clinical trial.

Three types of trials can be studied:

- Equivalence trials: An equivalence trial is where you want to demonstrate that a new treatment is no better or worse than an existing treatment.
- Superiority trials: A superiority trial is one where you want to demonstrate that one treatment is better than another.
- Non-inferiority trials: A non-inferiority trial is one where you want to show that a new treatment is not worse than an existing treatment.

These tests can be applied to a binary outcome or a continuous outcome.

When testing a hypothesis using a statistical test, there are several decisions to take:

- The null hypothesis H0 and the alternative hypothesis Ha.
- The statistical test to use.
- The type I error also known as alpha. It occurs when one rejects the null hypothesis when it is true. It is set a priori for each test and is 5%.

The type II error or beta is less studied but is of great importance. In fact, it represents the probability that one does not reject the null hypothesis when it is false. We can not fix it upfront, but based on other parameters of the model we can try to minimize it. The power of a test is calculated as $1 - beta$ and represents the probability that we reject the null hypothesis when it is false.

We therefore wish to maximize the power of the test. XLSTAT calculates the power (and beta) when other parameters are known. For a given power, it also allows to calculate the sample size that is necessary to reach that power.

The statistical power calculations are usually done before the experiment is conducted. The main application of power calculations is to estimate the number of observations necessary to properly conduct an experiment.

Methods

The necessary sample size is obtained using simple approximation methods.

Equivalence test for a continuous outcome

The mean outcome is compared between two randomised groups. You must define a difference between these means, d, within which you will accept that the two treatments being compared are equivalent.

The sample size is obtained using:

$$n = \frac{f(\alpha, \beta/2) \cdot 2 \cdot \sigma^2}{(d)^2}$$

With σ^2 being the variance of the outcome and:

$$f(\alpha, \beta) = (\Phi^{-1}(\alpha) + \Phi^{-1}(\beta))^2$$

Equivalence test for a binary outcome

The percentage of patients that "survived" is compared between two randomised groups. You must define a difference between these percentages, d, within which you will accept that the two treatments being compared are equivalent.

The sample size is obtained using:

$$n = \frac{f(\alpha, \beta/2) \cdot (P(std) \cdot (100 - P(std)))}{(P(std) - d)^2}$$

With $P(std)$ being the percentage for the treatments (we suppose these percentage are equivalent for both treatments), d is defined by the user and:

$$f(\alpha, \beta) = (\Phi^{-1}(\alpha) + \Phi^{-1}(\beta))^2$$

Non-inferiority test for a continuous outcome

The mean outcome is compared between two randomised groups. The null hypothesis is that the experimental treatment is inferior to the standard treatment. The alternative hypothesis is that the experimental treatment is non-inferior to the standard treatment. You must choose the

non-inferiority limit d, to be the largest difference that is clinically acceptable, so that a difference bigger than this would matter in practice.

The sample size is obtained using:

$$n = \frac{f(\alpha, \beta) \cdot 2 \cdot \sigma^2}{(d)^2}$$

With σ^2 being the variance of the outcome and:

$$f(\alpha, \beta) = (\Phi^{-1}(\alpha) + \Phi^{-1}(\beta))^2$$

Non-inferiority test for a binary outcome

The percentage of patients that "survived" is compared between two randomised groups. The null hypothesis is that the percentage for those on the standard treatment is better than the percentage for those on the experimental treatment by an amount d. The alternative hypothesis is that the experimental treatment is better than the standard treatment or only slightly worse (by no more than d). The user must define the non-inferiority limit (d) so that a difference bigger than this would matter in practice. You should normally assume that the percentage 'success' in both standard and experimental treatment groups is the same.

The sample size is obtained using:

$$n = \frac{f(\alpha, \beta) \cdot (P(std) \cdot (100 - P(std)) + P(new) \cdot (100 - P(new)))}{(P(std) - P(new) - d)^2}$$

With $P(std)$ being the percentage for the standard treatment and $P(new)$ being the percentage for the new treatment, d is defined by the user and:

$$f(\alpha, \beta) = (\Phi^{-1}(\alpha) + \Phi^{-1}(\beta))^2$$

Superiority test for a continuous outcome

The mean outcome is compared between two randomised groups. We wish to know if the mean associated to a new treatment is higher than the mean with the standard treatment.

The sample size is obtained using:

$$n = \frac{f(\alpha/2, \beta) \cdot 2 \cdot \sigma^2}{(\mu_1 - \mu_2)^2}$$

With σ^2 being the variance, μ_1 and μ_2 being the means associated to each group of the outcome and:

$$f(\alpha, \beta) = (\Phi^{-1}(\alpha) + \Phi^{-1}(\beta))^2$$

When cross-over is present, a formula for adjusting the sample size is used:

$$n_{\text{adjusted}} = \frac{n^* 10000}{(100 - c_1 - c_2)}$$

With c_1 and c_2 being the cross-over percentage in each group.

Superiority test for a binary outcome

The percentage of patients that "survived" is compared between two randomised groups. We wish to know if the percentage associated to a new treatment is higher than the percentage with the standard treatment.

The sample size is obtained using:

$$n = \frac{f(\alpha/2, \beta) \cdot (P(std) \cdot (100 - P(std)) + P(new) \cdot (100 - P(new)))}{(P(std) - P(new))^2}$$

With $P(std)$ being the percentage for the standard treatment and $P(new)$ being the percentage for the new treatment and:

$$f(\alpha, \beta) = (\Phi^{-1}(\alpha) + \Phi^{-1}(\beta))^2$$

When cross-over is present, a formula for adjusting the sample size is used:

$$n_{\text{adjusted}} = \frac{n^* 10000}{(100 - c_1 - c_2)}$$

With c_1 and c_2 being the cross-over percentage in each group.

Calculating power

To calculate the power for a fixed sample size, XLSTAT uses an algorithm that searches the beta (1-power) so that:

Sample size (beta) – expected sample size =0

We then obtain the power (1-beta) such that the test needs a sample size as close as possible to the desired sample size.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. Below are descriptions of the various elements of the dialog box.

OK: Click this button to start the calculations.

Cancel: Click this button to close the dialog box without doing any calculations.

Help: Click this button to display help options.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

General tab:

Goal: Choose between computing power and sample size estimation.

Clinical trial: Select the type of clinical trial: Equivalence, non- inferiority or superiority trials.

Outcome variable: Select the type of outcome variable (continuous or binary).

Alpha: Enter the value of the type I error (alpha, between 0.001 and 0.999).

Power (when sample size estimation has been selected): Enter the value of the power to be reached.

Sample size (when power computation has been selected): Enter the size of the total trial.

The available options will differ with respect to the chosen trial:

Equivalence trial with continuous outcome

Std deviation: Enter the standard deviation of the outcome.

Equivalence limit d: Enter the equivalence limit d.

Equivalence trial with binary outcome

% of success for both groups: Enter the % of success for both groups.

Equivalence limit d: Enter the equivalence limit d.

Non inferiority trial with continuous outcome

Std deviation: Enter the standard deviation of the outcome.

Non inferiority limit d: Enter the non inferiority limit d.

Non inferiority trial with binary outcome

% of success for control group: Enter the % of success for the control group.

% of success for treatment group: Enter the % of success for the treatment group.

Non inferiority limit d: Enter the non inferiority limit d.

Superiority trial with continuous outcome

Mean for control group: Enter the mean for the control group.

Mean for treatment group: Enter the mean for the treatment group.

Std deviation: Enter the standard deviation of the outcome.

% cross over for control group:: Enter the percentage of cross-over for the control group.

% cross over for treatment group:: Enter the percentage of cross-over for the treatment group.

Superiority trial with binary outcome

% of success for control group: Enter the % of success for the control group.

% of success for treatment group: Enter the % of success for the treatment group.

% cross over for control group:: Enter the percentage of cross-over for the control group.

% cross over for treatment group:: Enter the percentage of cross-over for the treatment group.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Graphics tab:

Simulation plot: Activate this option if you want to plot different parameters of the test. Two parameters can vary. All remaining parameters are used as they were defined in the General tab.

X axis: Select the parameter to be used on the X axis of the simulation plot. You can either choose the power or the sample size, the type I error (alpha) or the effect size. Depending on what we are looking for, we will have on the Y axis either the power or the sample size.

Interval size: Enter the minimum, maximum and interval size for the X axis of the simulation plot.

Results

Results: This table displays the parameters of the test and the power or the required number of observations. The parameters obtained by the calculation are in bold format. An explanation is displayed below this table.

Intervals for the simulation plot: This table is composed of two columns: power and sample size or alpha depending on the parameters selected in the dialog box. It helps building the simulation plot.

Simulation plot: This plot shows the evolution of the parameters as defined in the graphics tab of the dialog box.

Example

An example of calculating the required sample size for clinical trials is available on the XLSTAT Help Center at

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

References

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.

Statistical Process Control

Subgroup Charts

Use this tool to supervise production quality where you have a group of measurements for each point in time. The measurements need to be quantitative data. This tool is useful to recap the mean and the variability of the measured production quality.

This tool includes Integrated Box-Cox transformations, calculation of process capability and the application of rules for special causes and Westgard rules (an alternative set of rules to identify special causes) to complete your analyses.

In this section:

Description

Control charts were first mentioned in a document Walter Shewhart wrote during his time working at Bell Labs in 1924. He described his methods completely in his book (1931).

For a long time, there was no significant innovation in the area of control charts. With the development of CUSUM, UWMA and EWMA charts in 1936, Deming expanded the set of available control charts.

Control charts were originally used in factory production of goods. Therefore the wording is still from that domain. Today this approach is being applied in a large number of different fields, including services, human resources, and sales. In the following chapters, we will use the wording from the production and shop floors.

Subgroup charts

The subgroup charts tool offers you the following chart types alone or in combination:

- $\overline{X}$ (X bar): An X bar chart is useful to follow the mean of a production process. Mean shifts are easily visible in the diagrams.
- R: An R chart (Range chart) is useful to analyze the variability of the production. A large difference in production, caused for example by the use of different production lines, will be easily visible.

- S, S²: S and S² charts are also used to analyze the variability of production. The S chart draws the standard deviation of the process and the S² chart draws the variance (which is the square of the standard deviation).

Note 1: If you want to investigate smaller mean shifts, then you can also use CUSUM group charts which are, by the way, often preferred to subgroup control charts.

Note 2: If you have only one measurement for each point in time, then please use the control charts for individuals.

Note 3: If you have measurements in qualitative values (for instance ok, not ok, conform not conform), then use the control charts for attributes.

This tool offers you the following options for the estimation of the standard deviation (sigma or σ) of the data set, given k subgroups and $n_i (i = 1, \ldots, k)$ measurements per subgroup:

- Pooled standard deviation: sigma is computed using the k within-subgroup variances:

$$s = \sqrt{\frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)}} / c_4 \left(1 + \sum_{i=1}^{k}(n_i - 1)\right)$$

where c_4 is the control chart constant according to Burr (1969).

- R bar: The estimator for sigma is calculated based on the average range of the k subgroups.

$$s = \overline{R}/d_2$$

where d_2 is the control chart constant according to Burr (1969).

- S bar: The estimator for sigma is calculated based on the average of the standard deviations of the k subgroups:

$$s = \sqrt{\frac{1}{k}\sum_{i=1}^{k} s_i^2} / c_4$$

where c_4 is the control chart constant according to Burr (1969).

Process capability

Process capability describes a process and determines whether the process is under control and if values taken by the measured variables are inside the specification limits of the process. In the latter case, one says that the process is "capable".

During the interpretation of the different indicators of process capability, please pay attention to the fact that some indicators suppose normality or at least symmetry of the distribution of the measured values. Through the use of a normality test, you can verify these premises (see the Normality Tests in XLSTAT).

If the data are not normally distributed, you have the following possibilities to obtain results for the process capabilities.

- Use the Box-Cox transformation to improve the normality of the data set. Then verify again the normality using a normality test.
- Use the process capability indicator Cp 5.15.

Let m, s, LSL, USL be respectively the estimated mean, standard deviation, lower specification limit, upper specification limit of the process, and T be the selected target. Let μ and σ be the true mean and standard deviation of the process. XLSTAT allows to compute the following performance indicators to evaluate the process capability:

- Cp: The short term process capability is defined as:

$$Cp = \frac{USL - LSL}{6s}$$

- Cpl: The short term process capability with respect to the lower specification is defined as:

$$Cpl = \frac{m - LSL}{3s}$$

- Cpu: The short term process capability with respect to the upper specification is defined as:

$$Cpu = \frac{USL - m}{3s}$$

- Cpk: The short term process capability supposing a centered distribution is defined as:

$$Cpk = min(Cpl, Cpu)$$

- Pp: The long term process capability is defined as:

$$Pp = \frac{USL - LSL}{6\sigma}$$

- Ppl: The long term process capability with respect to the lower specification is defined as:

$$Ppl = \frac{m - LSL}{3\sigma}$$

- Ppu: The long term process capability with respect to the upper specification is defined as:

$$Ppu = \frac{USL - m}{3\sigma}$$

- Ppk: The long term process capability supposing a centered distribution is defined as:

$$Ppk = min(Ppl, Ppu)$$

- Cpm: The short term process capability according to Taguchi. This value can be calculated, if the target value has been specified. It is defined as:

$$Cpm = \frac{USL - LSL}{6\sqrt{s^2 + (m - T)^2}}$$

- Cpm Boyles: The short term process capability according to Taguchi improved by Boyles (1991). This value can be calculated if the target value has been specified. It is defined as:

$$Cpm\ Boyles = \frac{USL - LSL}{6\sqrt{(n-1)s^2/n + (m - T)^2}}$$

- Cp 5.15: The short term process capability is defined as:

$$Cp\ 5.15 = \frac{USL - LSL}{5.15s}$$

- Cpk 5.15: The short term process capability supposing a centered distribution is defined as:

$$Cpk\ 5.15 = \frac{(USL - LSL)/2 - |m - (USL + LSL)/2|}{2.57s}$$

- Cpmk: The short term process capability according to Pearn. This value can be calculated if the target value has been specified. It is defined as:

$$Cpmk = \frac{s \cdot Cpk}{\sqrt{s^2 + (m - T)^2}}$$

- Cs Wright: The process capability according to Wright. This value can be calculated if the target value has been specified. It is defined as:

$$Cs = \frac{(USL - LSL)/2 - |m - T|}{3\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - T)^2 + \left|\frac{n^2 m_3}{(n-1)(n-2)} \times \left(\frac{n}{n-1} \times \frac{m_2}{c_4^2}\right)^{-1/2}\right|}}$$

with

$$m_r = \sum_{i=1}^{n}(x_i - T)^r$$

$$c_4 = \left[\frac{2}{n-1}\right]^{1/2} \Gamma(\frac{n}{2})/\Gamma(\frac{n-1}{2})^{-1}$$

- Z below: The amount of standard deviations between the mean and the lower specification limit is defined as:

$$Zbelow = (m - LSL)/s$$

- Z above: The amount of standard deviations between the mean and the upper specification limit is defined as:

$$Zabove = (USL - m)/s$$

- Z total: The amount of standard deviations between the mean and the lower or upper specification limit is defined as:

$$Ztotal = \Phi^{-1}(p(\text{not conform})total)$$

- p(not conform) below: The probability of producing a defect product below the lower specification limit is defined as:

$$p(\text{not conform})below = \Phi(Zbelow)$$

- p(not conform) above: The probability of producing a defect product above the upper specification limit is defined as:

$$p(\text{not conform})above = \Phi(Zabove)$$

- p(not conform) total: The probability of producing a defect product below or above the specification limits is defined as:

$$p(\text{not conform})total = p(\text{not conform})below + p(\text{not conform})above$$

- PPM below: The number of defective products below the lower specification limit over one million items produced is defined as:

$$PPMbelow = p(notconform)below \times 10^6$$

- PPM above: The number of defective products above the upper specification limit over one million items produced is defined as:

$$PPMabove = p(notconform)above \times 10^6$$

- PPM total: The number of defective products below or above the specification limits over one million items produced is defined as:

$$PPMtotal = PPMbelow + PPMabove$$

Box-Cox transformation

Box-Cox transformation is used to improve the normality of the time series; the Box-Cox transformation is defined by the following equation:

$$Y_t = \begin{cases} \frac{X_t^\lambda - 1}{\lambda}, & X_t \geq 0, \lambda > 0 \\ \ln(X_t), & X_t > 0, \lambda = 0 \end{cases}$$

Where the series $\{X_t\}$ is being transformed into series $\{Y_t\}$, (t=1,…,n):

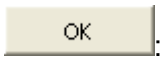
Note: if $\lambda < 0$, the first equation is still valid, but Xt must be strictly positive. XLSTAT accepts a fixed value of λ, or it can find the value that maximizes the likelihood value, the model being a simple linear model with the time as sole explanatory variable.

Chart rules

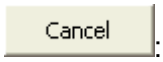
XLSTAT lets you apply rules for special causes and Westgard rules. Two sets of rules are available in order to interpret control charts. You can activate and deactivate the rules in each set separately.

Dialog box

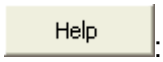
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below descriptions of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help menu.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files  .

Mode tab:

Chart family: Select the family that you want to use:

- **Subgroup charts**: Activate this option if you have a data set with several measurements for each point in time.

- **Individual charts**: Activate this option if you have a data set with one quantitative measurement for each point in time.
- **Attribute charts**: Activate this option if you have a data set with one qualitative measurement for each point.
- **Time weighted**: Activate this option if you want to use a time weighted chart like UWMA, EWMA or CUSUM.

At this stage, the subgroup charts family should be selected. If not, you should switch to the help corresponding to the selected chart family. The options below correspond to the subgroup charts

Chart type: Select the type of chart you want to use:

- **X bar chart**: Activate this option if you want to calculate the X bar chart to analyze the mean of the process.
- **R chart**: Activate this option if you want to calculate the R chart to analyze variability of the process.
- **S chart**: Activate this option if you want to calculate the S chart to analyze variability of the process.
- **S² chart**: Activate this option if you want to calculate the S² chart to analyze variability of the process.
- **X bar R chart**: Activate this option if you want to calculate the X bar chart together with the R chart to analyze the mean value and variability of the process.
- **X bar S chart**: Activate this option if you want to calculate the X bar chart together with the S chart to analyze the mean value and variability of the process.
- **X bar S² chart**: Activate this option if you want to calculate the X bar chart together with the S² chart to analyze the mean value and variability of the process.

General tab:

Data format: Select the data format.

- **Columns/Rows**: Activate this option for XLSTAT to take each column (in column mode) or each row (in row mode) as a separate measurement that belongs to the same subgroup.
- **One column/row**: Activate this option if the measurements of the different subgroups are all on the same column (column mode) or one row (row mode). To assign the different measurements to their corresponding subgroup, please enter a constant group size or select a column or row with the group identifier in it.

Data: If the data format "One column/row" is selected, please choose the unique column or row that contains all the data. The assignment of the data to their corresponding subgroup must be specified using the Groups field or by setting the common subgroup size. If you select the data "Columns/rows" option, please select a data area with one column/row per measurement in a subgroup.

Groups: If the data format "One column/row" is selected, then activate this Option to select a column/row that contains the group identifier. Select the data that identify the corresponding group for each element of the data selection .

Common subgroup size: If the data format "One column/row" is selected and the subgroup size is constant, then you can deactivate the groups option and enter in this field the common subgroup size.

Phase: Activate this option to supply one column/row with the phase identifier. At each phase, XLSTAT will recalculate the central line and the control limits and create a new chart.

- **Different specifications**: Activate this option if you want to enter specifications specific to each phase for the process capability parameters. In this case, in the Options tab, you must enter a USL value, an LSL value and a target value for each phase.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row (column mode) or column (row mode) of the data selections contains a label.

Options tab:

Upper control limit:

- **Bound**: Activate this option, if you want to enter a maximum value to accept for the upper control limit of the process. This value will be used when the calculated upper control limit is greater than the value entered here.
- **Value**: Enter the upper control limit. This value will be used in place of the calculated upper control limit.

Lower control limit:

- **Bound**: Activate this option if you want to enter a minimum value to accept for the lower control limit of the process. This value will be used when the calculated lower control limit is greater than the value entered here.
- **Value**: Enter the lower control limit. This value will be used and overrides the calculated upper control limit.

Calculate process capabilities: Activate this option to calculate process capabilities based on the input data (see the description section for more details).

- **USL**: If the calculation of the process capabilities is activated, please enter the upper specification limit (USL) of the process here.
- **LSL**: If the calculation of the process capabilities is activated, please enter the lower specification limit (LSL) of the process here.
- **Target**: If the calculation of the process capabilities is activated, activate this option to add the target value of the process.
- **Confidence interval (%)**: If the "Calculate process capabilities" option is activated, please enter the percentage range of the confidence interval to use for calculating the confidence interval around the process capabilities. Default value: 95.

Box-Cox: Activate this option to compute the Box-Cox transformation. You can either fix the value of the Lambda parameter, or decide to let *XL* STAT optimize it (see the description for further details).

k Sigma: Activate this option to enter the distance between the upper and the lower control limit and the center line of the control chart. The distance is fixed to k times the factor you enter multiplied by the estimated standard deviation. Corrective factors according to Burr (1969) will be applied.

alpha: Activate this option to define the size of the confidence range around the center line of the control chart. 100 - alpha % of the distribution of the control chart is inside the control limits. Corrective factors according to Burr (1969) will be applied.

Mean: Activate this option to enter a value for the center line of the control chart. This value should be based on historical data.

Sigma: Activate this option to enter a value for the standard deviation of the control chart. This value should be based on historical data. If this option is activated, then you cannot choose an estimation method for the standard deviation in the "Estimation" tab.

Estimation tab:

Method for Sigma: Select an option to determine the estimation method for the standard deviation of the control chart (see the description for further details):

- Pooled standard deviation
- R-bar
- S-bar

Outputs tab:

Display zones: Activate this option to display beside the lower and upper control limit also the limits of the zones A and B.

Normality Tests: Activate this option to check normality of the data. (see the Normality Tests tool for further details).

Significance level (%): Enter the significance level for the tests.

Test special causes: Activate this option to analyze the points of the control chart according to the rules for special causes. You can activate the following rules independently:

- 1 point more than 3s from center line
- 9 points in a row on same side of center line
- 6 points in a row, all increasing or all decreasing
- 14 points in a row, alternating up and down
- 2 out of 3 points > 2s from center line (same side)
- 4 out of 5 points > 1s from center line (same side)
- 15 points in a row within 1s of center line (either side)
- 8 points in a row > 1s from center line (either side)
- **All**: Click this button to select all options.
- **None**: Click this button to deselect all options.

Apply Westgard rules: Activate this option to analyze the points of the control chart according to the Westgard rules. You can activate the following rules independently:

- Rule 1 2s
- Rule 1 3
- Rule 2 2s

- Rule 4s
- Rule 4 1s
- Rule 10 X
- **All**: Click this button to select all options.
- **None**: Click this button to deselect all options.

Charts tab:

Display charts: Activate this option to display the control charts graphically.

Normal Q-Q plots: Check this option to display Q-Q plots based on the normal distribution.

Histograms: Activate this option to display the histograms of the samples. For a theoretical distribution, the density function is displayed.

Run Charts: Activate this option to display a chart of the latest data points. Each individual measurement is displayed.

Results

Estimation:

Estimated mean: This table displays the estimated mean values for the different phases.

Estimated standard deviation: This table displays the estimated standard deviation values for the different phases.

Box-Cox transformation:

Estimates of the parameters of the model: This table is available only if the Lambda parameter has been optimized. It displays the estimator for Lambda.

Series before and after transformation: This table displays the series before and after transformation. If Lambda has been optimized, the transformed series corresponds to the residuals of the model. If it hasn't then the transformed series is the direct application of the Box-Cox transformation

Process capabilities:

Process capabilities: These tables are displayed, if the "process capability" option has been selected. There is one table for each phase. A table contains the following indicators for the process capability and if possible the corresponding confidence intervals: Cp, Cpl, Cpu, Cpk, Pp, Ppl, Ppu, Ppk, Cpm, Cpm (Boyle), Cp 5.5, Cpk 5.5, Cpmk, and Cs (Wright).

For Cp, Cpl, and Cpu, information about the process performance is supplied and for Cp a status information is given to facilitate the interpretation.

Cp values have the following status based on Ekvall and Juran (1974):

- "not adequate" if $Cp < 1$
- "adequate" if $1 \leq Cp \leq 1.33$
- "more than adequate" if $Cp > 1.33$

Based on Montgomery (2001), Cp needs to have the following minimal values for the process performance to be as expected:

- 1.33 for existing processes
- 1.50 for new processes or for existing processes when the variable is critical
- 1.67 for new processes when the variable is critical

Based on Montgomery (2001), Cpu and Cpl need to have the following minimal values for process performance to be as expected:

- 1.25 for existing processes
- 1.45 for new processes or for existing processes when the variable is critical
- 1.60 for new processes when the variable is critical

Capabilities: This chart contains information about the specification and control limits. A line between the lower und upper limits represents the interval with an additional vertical mark for the center line. The different control limits of each phase are drawn separately.

Charts informations :

The following results are displayed separately for each requested chart. Charts can be selected alone or in combination with the X bar chart.

X bar/ R/ S/ S² chart: This table contains information about the center line and the upper and lower control limits of the selected chart. There will be one column for each phase.

Observation details: This table displays detailed information for each subgroup. For each subgroup, the corresponding phase, the size, the mean, the minimum and the maximum values, the center line, and the lower and upper control limits are displayed. If the information about the zones A, B and C are activated, then the lower and upper control limits of the zones A and B are displayed as well.

Rule details: If the rules options are activated, a detailed table about the rules will be displayed. For each subgroup, there is one row for each rule that applies. "Yes" indicates that the corresponding rule was fired for the corresponding subgroup and "No" indicates that the rule does not apply.

X bar/ R/ S/ S² chart: If the charts are activated, then a chart containing the information of the two tables above is displayed. Each subgroup is displayed. The center line and the lower and upper control limits are displayed as well. If the corresponding options have been activated, the lower and upper control limits for the zones A and B are included and there are labels for the subgroups for which rules were fired. A legend with the activated rules and the corresponding rule number is displayed below the chart.

Normality tests :

For each of the four tests, the statistics relating to the test are displayed including, in particular, the p-value, which is later used in interpreting the test by comparing with the chosen significance threshold.

If requested, a Q-Q plot is then displayed.

Histograms: The histograms are displayed. If desired, you can change the color of the lines, scales, titles as with any Excel chart.

Run chart: The chart of the last data points is displayed.

Example

A tutorial explaining how to use the SPC subgroup charts tool is available on the XLSTAT Help Center. To consult the tutorial, please go to:

http://www.xlstat.com/demoSPS_EN.htm

References

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.

Individual Charts

Use this tool to supervise production quality where you have a single measurement for each point in time. The measurements need to be quantitative variables.

This tool is useful to recap the moving average and median and the variability of the production quality that is being measured.

This tool includes Integrated Box-Cox transformations, calculation of process capability and the application of rules for special causes and Westgard rules (an alternative rule set to identify special causes) available to complete your analysis.

In this section:

Description

Control charts were first mentioned in a document Walter Shewhart wrote during his time working at Bell Labs in 1924. He described his methods completely in his book (1931).

For a long time, there was no significant innovation in the area of control charts. With the development of CUSUM, UWMA and EWMA charts in 1936, Deming expanded the set of available control charts.

Control charts were originally used in factory production of goods. Therefore the wording is still from that domain. Today this approach is being applied to a large number of different fields, including services, human resources, and sales. In the following lines, we use the wording from the production and shop floors.

Individual charts

The individual charts tool offers you the following chart types, alone or in combination:

- X Individual: It is useful to follow the moving average of a production process. Mean shifts are easily visible in the diagrams.
- MR moving range: It is useful to analyze the variability of production. Large differences in production, caused by the use of different production lines, will be easily visible.

Note 1: If you want to investigate smaller mean shifts, then you can also use CUSUM individual charts which are often preferred in comparison with the individual control charts, because they

can detect smaller mean shifts.

Note 2: If you have more than one measurement for each point in time, then you should use the control charts for subgroups.

Note 3: If you have measurements in qualitative values (for instance ok, not ok, conform not conform), then use the control charts for attributes.

This tool offers you the following options for the estimation of the standard deviation (sigma) of the data set, given n measurements:

- Average moving range: The estimator for sigma is calculated based on the average moving range using a window length of m measurements.

$$\hat{s} = \overline{m}/d_2$$

where d_2 is the control chart constant according to Burr (1969).

- Median moving range: The estimator for sigma is calculated based on the median of the moving range using a window length of m measurements.

$$\hat{s} = \overline{median}/d_4$$

where d_4 is the control chart constant according to Burr (1969).

- standard deviation: The estimator for sigma is calculated based on the standard deviation s of the n measurements.

$$\hat{s} = s/c_4$$

where c_4 is the control chart constant according to Burr (1969).

Process capability

Process capability describes a process and determines whether the process is under control and the distribution of the measured variables are inside the specification limits of the process. If the distributions of the measured variables are in the technical specification limits, then the process is called "capable".

During the interpretation of the different indicators for process capability, please pay attention to the fact that some indicators suppose normality or at least symmetry of the distribution of the measured values. By using a normality test, you can verify these premises (see the Normality Tests in XLSTAT-Pro).

If the data are not normally distributed, you have the following possibilities to obtain results for the process capabilities.

- Use the Box-Cox transformation to improve the normality of the data set. Then verify again the normality using a normality test.

- Use the process capability indicator Cp 5.15.

Let m, s, LSL, USL be (respectively) the estimated mean, standard deviation, lower specification limit, upper specification limit of the process, and T be the selected target. Let μ and σ be the true mean and standard deviation of the process. XLSTAT lets you compute the following performance indicators to evaluate the process capability:

- Cp: The short term process capability is defined as:

$$Cp = \frac{USL - LSL}{6s}$$

- Cpl: The short term process capability with respect to the lower specification is defined as:

$$Cpl = \frac{m - LSL}{3s}$$

- Cpu: The short term process capability with respect to the upper specification is defined as:

$$Cpu = \frac{USL - m}{3s}$$

- Cpk: The short term process capability supposing a centered distribution is defined as:

$$Cpk = min(Cpl, Cpu)$$

- Pp: The long term process capability is defined as:

$$Pp = \frac{USL - LSL}{6\sigma}$$

- Ppl: The long term process capability with respect to the lower specification is defined as:

$$Ppl = \frac{m - LSL}{3\sigma}$$

- Ppu: The long term process capability with respect to the upper specification is defined as:

$$Ppu = \frac{USL - m}{3\sigma}$$

- Ppk: The long term process capability supposing a centered distribution is defined as:

$$Ppk = min(Ppl, Ppu)$$

- Cpm: The short term process capability according to Taguchi. This value can be calculated, if the target value has been specified. It is defined as:

$$Cpm = \frac{USL - LSL}{6\sqrt{s^2 + (m - T)^2}}$$

- Cpm Boyles: The short term process capability according to Taguchi improved by Boyles (1991). This value can be calculated if the target value has been specified. It is defined as:

$$Cpm\ Boyles = \frac{USL - LSL}{6\sqrt{(n-1)s^2/n + (m-T)^2}}$$

- Cp 5.15: The short term process capability is defined as:

$$Cp\ 5.15 = \frac{USL - LSL}{5.15s}$$

- Cpk 5.15: The short term process capability supposing a centered distribution is defined as:

$$Cpk\ 5.15 = \frac{(USL - LSL)/2 - |m - (USL + LSL)/2|}{2.57s}$$

- Cpmk: The short term process capability according to Pearn. This value can be calculated if the target value has been specified. It is defined as:

$$Cpmk = \frac{s \cdot Cpk}{\sqrt{s^2 + (m-T)^2}}$$

- Cs Wright: The process capability according to Wright. This value can be calculated if the target value has been specified. It is defined as:

$$Cs = \frac{(USL - LSL)/2 - |m - T|}{3\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - T)^2 + \left|\frac{n^2 m_3}{(n-1)(n-2)} \times \left(\frac{n}{n-1} \times \frac{m_2}{c_4^2}\right)^{-1/2}\right|}}$$

with

$$m_r = \sum_{i=1}^{n}(x_i - T)^r$$

$$c_4 = \left[\frac{2}{n-1}\right]^{1/2} \Gamma(\frac{n}{2})/\Gamma(\frac{n-1}{2})^{-1}$$

- Z below: The amount of standard deviations between the mean and the lower specification limit is defined as:

$$Zbelow = (m - LSL)/s$$

- Z above: The amount of standard deviations between the mean and the upper specification limit is defined as:

$$Zabove = (USL - m)/s$$

- Z total: The amount of standard deviations between the mean and the lower or upper respectively specification limit is defined as:

$$Ztotal = \Phi^{-1}(p(\text{not conform})total)$$

- p(not conform) below: The probability of producing a defective product below the lower specification limit is defined as:

$$p(\text{not conform})below = \Phi(Zbelow)$$

- p(not konform) above: The probability of producing a defective product above the upper specification limit is defined as:

$$p(\text{not conform})above = \Phi(Zabove)$$

- p(not conform) total: The probability of producing a defective product below or above the specification limits is defined as:

$$p(\text{not conform})total = p(\text{not conform})below + p(\text{not conform})above$$

- PPM below: The number of defective products below the lower specification limit over one million items produced is defined as:

$$PPMbelow = p(notconform)below \times 10^6$$

- PPM above: The number of defective products above the upper specification limit over one million items produced is defined as:

$$PPMabove = p(notconform)above \times 10^6$$

- PPM total: The number of defective products below or above the specification limits over one million items produced is defined as:

$$PPMtotal = PPMbelow + PPMabove$$

Box-Cox transformation

Box-Cox transformation is used to improve the normality of the time series; the Box-Cox transformation is defined by the following equation:

$$Y_t = \begin{cases} \frac{X_t^\lambda - 1}{\lambda}, & X_t \geq 0, \lambda > 0 \\ \ln(X_t), & X_t > 0, \lambda = 0 \end{cases}$$

Where the series $\{X_t\}$ is being transformed into series $\{Y_t\}$, (t=1,…,n):

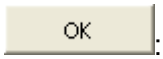
Note: if $\lambda < 0$ the first equation is still valid, but X_t must be strictly positive. XLSTAT accepts a fixed value of λ, or it can find the value that maximizes the likelihood value, the model being a simple linear model with the time as sole explanatory variable.

Chart rules

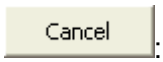
XLSTAT lets you apply rules for special causes and Westgard rules on the data set. Two sets of rules are available in order to interpret control charts. You can activate and deactivate the rules in each set separately.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. Below are descriptions of the various elements of the dialog box.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help options.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Chart type: Select the type of chart you want to use:

- **X Individual chart**: Activate this option if you want to calculate the X individual chart to analyze the mean of the process.
- **MR Moving Range chart**: Activate this option if you want to calculate the MR chart to analyze variability of the process.
- **X-MR Individual/Moving Range chart**: Activate this option if you want to calculate the X Individual chart together with the MR chart to analyze the mean value and variability of the process.

Data: Please choose the unique column or row that contains all the data.

Phase: Activate this option to supply one column/row with the phase identifier.

- **Different specifications**: Activate this option if you want to enter specifications specific to each phase for the process capability parameters. In this case, in the Options tab, you must enter a USL value, an LSL value and a target value for each phase.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated, you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row (column mode) or column (row mode) of the data selections contains a label.

Options tab:

Upper control limit:

Bound: Activate this option if you want to enter a maximum value to accept for the upper control limit of the process. This value will be used when the calculated upper control limit is greater than the value entered here.

Value: Enter the upper control limit. This value will be used and overrides the calculated upper control limit.

Lower control limit:

Bound: Activate this option, if you want to enter a minimum value to accept for the lower control limit of the process. This value will be used when the calculated lower control limit is greater than the value entered here.

Value: Enter the lower control limit. This value will be used in place of the calculated upper control limit.

Calculate Process capabilities: Activate this option to calculate process capabilities based on the input data (see the description section for more details).

USL: If the calculation of the process capabilities is activated, please enter the upper specification limit (USL) of the process here.

LSL: If the calculation of the process capabilities is activated, please enter the lower specification limit (LSL) of the process here.

Target: If the calculation of the process capabilities is activated, activate this option to add the target value of the process.

Confidence interval (%):If the "Calculate Process Capabilities" option is activated, please enter the percentage range of the confidence interval to use for calculating the confidence interval around the parameters. Default value: 95.

Box-Cox: Activate this option to compute the Box-Cox transformation. You can either fix the value of the Lambda parameter, or decide to let *XL* STAT optimize it (see the description for further details).

k Sigma: Activate this option to enter the distance between the upper and the lower control limit and the center line of the control chart. The distance is fixed to k times the factor you enter, multiplied by the estimated standard deviation. Corrective factors according to Burr (1969) will be applied.

alpha: Activate this option to enter the size of the confidence range around the center line of the control chart. The alpha is used to compute the upper and lower control limits. 100 – alpha % of the distribution of the control chart is inside the control limits. Corrective factors according to Burr (1969) will be applied.

Mean: Activate this option to enter a value for the center line of the control chart. This value should be based on historical data.

Sigma: Activate this option to enter a value for the standard deviation of the control chart. This value should be based on historical data. If this option is activated, then you cannot choose an estimation method for the standard deviation in the "Estimation" tab.

Estimation tab:

Method for Sigma: Select an option to determine the estimation method for the standard deviation of the control chart (see the description for further details):

- Average Moving Range
- Median Moving Range
- MR Length: Change this value to modify the number of observations that are taken into account in the moving range.
- Standard deviation: The estimator of sigma is calculated using the standard deviation of the n measurements.

Outputs tab:

Display zones: Activate this option to display the limits of the zones A and B beside the lower and upper control limit as well.

Normality Tests: Activate this option to check normality of the data. (see the Normality Tests tool for further details).

Significance level (%): Enter the significance level for the tests.

Test special causes: Activate this option to analyze the points of the control chart according to the rules for special causes. You can activate the following rules independently:

- 1 point more than 3s from center line
- 9 points in a row on same side of center line
- 6 points in a row, all increasing or all decreasing
- 14 points in a row, alternating up and down
- 2 out of 3 points > 2s from center line (same side)
- 4 out of 5 points > 1s from center line (same side)
- 15 points in a row within 1s of center line (either side)
- 8 points in a row > 1s from center line (either side)
- **All**: Click this button to select all.
- **None**: Click this button to deselect all.

Apply Westgard rules: Activate this option to analyze the points of the control chart according to the Westgard rules. You can activate the following rules independently:

- Rule 1 2s
- Rule 1 3
- Rule 2 2s
- Rule 4s
- Rule 4 1s
- Rule 10 X
- **All**: Click this button to select all.
- **None**: Click this button to deselect all.

Charts tab:

Display charts: Activate this option to display the control charts graphically.

Normal Q-Q Charts: Check this option to display Q-Q plots.

Histograms: Activate this option to display the histograms of the samples. For a theoretical distribution, the density function is displayed.

Run Charts: Activate this option to display a chart of the latest data points. Each individual measurement is displayed.

Results

Estimation:

Estimated mean: This table displays the estimated mean values for the different phases.

Estimated standard deviation: This table displays the estimated standard deviation values for the different phases.

Box-Cox transformation:

Estimates of the parameters of the model: This table is available only if the Lambda parameter has been optimized. It displays the estimator for Lambda.

Series before and after transformation: This table displays the series before and after transformation. If Lambda has been optimized, the transformed series corresponds to the residuals of the model. If it hasn't then the transformed series is the direct application of the Box-Cox transformation.

Process capability:

Process capabilities: These tables are displayed, if the "process capability" option has been selected. There is one table for each phase. A table contains the following indicators for the process capability and if possible the corresponding confidence intervals: Cp, Cpl, Cpu, Cpk, Pp, Ppl, Ppu, Ppk, Cpm, Cpm (Boyle), Cp 5.5, Cpk 5.5, Cpmk, and Cs (Wright).

For Cp, Cpl, and Cpu, information about the process performance is supplied and for Cp a status information is given to facilitate the interpretation.

Cp values have the following status based on Ekvall and Juran (1974):

- "not adequate" if $Cp < 1$
- "adequate" if $1 \leq Cp \leq 1.33$
- "more than adequate" if $Cp > 1.33$

Based on Montgomery (2001), Cp needs to have the following minimal values for the process performance to be as expected:

- 1.33 for existing processes
- 1.50 for new processes or for existing processes when the variable is critical
- 1.67 for new processes when the variable is critical

Based on Montgomery (2001), Cpu and Cpl need to have the following minimal values for process performance to be as expected:

- 1.25 for existing processes
- 1.45 for new processes or for existing processes when the variable is critical
- 1.60 for new processes when the variable is critical

Capabilities: This chart contains information about the specification and control limits. A line between the lower und upper limits represents the interval with an additional vertical mark for the center line. The different control limits of each phase are drawn separately.

Chart information:

The following results are displayed separately for each requested chart. Charts can be selected alone or in combination with the X individual chart.

X Individual / MR moving range chart: This table contains information about the center line and the upper and lower control limits of the selected chart. There will be one column for each phase.

Observation details: This table displays detailed information for each observation. For each observation, the corresponding phase, the mean or median, the center line, the lower and upper control limits are displayed. If the information about the zones A, B and C are activated, then the lower and upper control limits of the zones A and B are displayed as well.

Rule details: If the rules options are activated, a detailed table about the rules will be displayed. For each observation, there is one row for each rule that applies. "Yes" indicates that the corresponding rule was fired, and "No" indicates that the rule does not apply.

X Individual / MR moving range Chart: If the charts are activated, then a chart containing the information of the two tables above is displayed. Each observation is displayed. The center line and the lower and upper control limits are displayed as well. If the corresponding options have been activated, the lower and upper control limits for the zones A and B are included and there are labels for the observations for which rules were activated. A legend with the activated rules and the corresponding rule number is displayed below the chart.

Normality tests:

For each of the four tests, the statistics relating to the test are displayed including, in particular, the p-value which is used afterwards in interpreting the test by comparing with the chosen significance threshold.

If requested, a Q-Q plot is then displayed.

Histograms: The histograms are displayed. If desired, you can change the color of the lines, scales, titles as with any Excel chart.

Run chart: The chart of the last data points is displayed.

Example

A tutorial explaining how to use the SPC subgroup charts tool is available on the XLSTAT Help Center at:

http://www.xlstat.com/demoSPI_EN.htm

References

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.

Attribute charts

Use this tool to supervise production quality, in the case where you have a single measurement for each point in time. The measurements are based on attributes or attribute counts of the process.

This tool is useful to recap the categorical variables of the measured production quality.

Integrated in this tool, you will find Box-Cox transformations, calculation of process capability and the application of rules for special causes and Westgard rules (an alternative rule set to identify special causes) available to complete your analyses.

In this section:

Description

Control charts were first mentioned in a document that Walter Shewhart wrote during his time working at Bell Labs in 1924. He described his methods completely in his book (1931).

For a long time, there had been no significant innovation in the area of control charts. With the development of CUSUM, UWMA and EWMA charts in 1936, Deming expanded the set of available control charts.

Control charts were originally used in factory production of goods. Therefore the wording is still from that domain. Today this approach is being applied to a large number of different fields, such as services, human resources, and sales. In the following chapters we will use the wording from production and shop floors.

Attribute charts

The attribute charts tool offers you the following chart types:

- P chart: This is useful to follow the fraction of non conforming units of a production process.
- NP chart: This is useful to follow the absolute number of non conforming units of a production process.
- C chart: This is useful in the case of a production having a constant size for each inspection unit. It can be used to follow the absolute number of non-conforming items per

inspection.

- U chart: This is useful in cases where the size of each inspection unit in production is not constant. It can be used to follow the fraction of the non conforming items per inspection.

These charts analyze either "nonconforming products" or "nonconformities." They are usually used to inspect quality before delivery (outgoing products) or quality at delivery (incoming products). Not all the products need necessarily be inspected.

Inspections are done by inspection units of well defined size. The size can be 1 in the case of the reception of television sets at a warehouse. The size would be 24 in the case of peaches delivered in crates of 24 peaches.

P and NP charts allow to analyze the fraction respectively the absolute number of nonconforming products of a production process. For example, we can count the number of nonconforming television sets, or the number of crates that contain at least one bruised peach.

C and U charts analyze the fraction or the absolute number, respectively, of occurrences of nonconformities in an inspection unit. For example, we can count the number of defective transistors for each inspection unit (there might be more than one transistor not working in one television set), or the number of bruised peaches per crate.

Process capability

Process capability describes a process and determines whether the process is under control and the distribution of the measured variables are inside the specification limits of the process. If the distributions of the measured variables are in the technical specification limits, then the process is called "capable."

While interpreting the different indicators for the process capability, please pay attention to the fact that some indicators suppose normality or at least symmetry of the distribution of the measured values. By using a normality test, you can verify these premises (see the Normality Tests in XLSTAT-Pro).

If the data are not normally distributed, you can use the following possibilities to obtain results for the process capabilities:

Use the Box-Cox transformation to improve the normality of the data set. Then re-verify the normality using a normality test.

Use the process capability indicator Cp 5.5.

Let m, s, LSL, USL be respectively the estimated mean, standard deviation, lower specification limit, upper specification limit of the process, and let T be the selected target. Let μ and σ be the true mean and standard deviation of the process. XLSTAT allows you to compute the following performance indicators to evaluate the process capability:

- Cp: The short-term process capability is defined as:

$$Cp = \frac{USL - LSL}{6s}$$

- Cpl: The short-term process capability with respect to the lower specification is defined as:

$$Cpl = \frac{m - LSL}{3s}$$

- Cpu: The short term process capability with respect to the upper specification is defined as:

$$Cpu = \frac{USL - m}{3s}$$

- Cpk: The short term process capability supposing a centered distribution is defined as:

$$Cpk = min(Cpl, Cpu)$$

- Pp: The long term process capability is defined as:

$$Pp = \frac{USL - LSL}{6\sigma}$$

- Ppl: The long term process capability with respect to the lower specification is defined as:

$$Ppl = \frac{m - LSL}{3\sigma}$$

- Ppu: The long term process capability with respect to the upper specification is defined as:

$$Ppu = \frac{USL - m}{3\sigma}$$

- Ppk: The long term process capability supposing a centered distribution is defined as:

$$Ppk = min(Ppl, Ppu)$$

- Cpm: The short term process capability according to Taguchi. This value can be calculated, if the target value has been specified. It is defined as:

$$Cpm = \frac{USL - LSL}{6\sqrt{s^2 + (m - T)^2}}$$

- Cpm Boyles: The short term process capability according to Taguchi, improved by Boyles (1991). This value can be calculated if the target value has been specified. It is defined as:

$$Cpm\ Boyles = \frac{USL - LSL}{6\sqrt{(n-1)s^2/n + (m - T)^2}}$$

- Cp 5.15: The short term process capability is defined as:

$$Cp\ 5.15 = \frac{USL - LSL}{5.15s}$$

- Cpk 5.15: The short term process capability supposing a centered distribution is defined as:

$$Cpk\ 5.15 = \frac{(USL - LSL)/2 - |m - (USL + LSL)/2|}{2.57s}$$

- Cpmk: The short term process capability according to Pearn. This value can be calculated if the target value has been specified. It is defined as:

$$Cpmk = \frac{s \cdot Cpk}{\sqrt{s^2 + (m - T)^2}}$$

- Cs Wright: The process capability according to Wright. This value can be calculated if the target value has been specified. It is defined as:

$$Cs = \frac{(USL - LSL)/2 - |m - T|}{3\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - T)^2 + \left|\frac{n^2 m_3}{(n-1)(n-2)} \times \left(\frac{n}{n-1} \times \frac{m_2}{c_4^2}\right)^{-1/2}\right|}}$$

with

$$m_r = \sum_{i=1}^{n}(x_i - T)^r$$

$$c_4 = \left[\frac{2}{n-1}\right]^{1/2} \Gamma(\frac{n}{2})/\Gamma(\frac{n-1}{2})^{-1}$$

- Z below: The amount of standard deviations between the mean and the lower specification limit is defined as:

$$Zbelow = (m - LSL)/s$$

- Z above: The amount of standard deviations between the mean and the upper specification limit is defined as:

$$Zabove = (USL - m)/s$$

- Z total: The amount of standard deviations between the mean and the lower or upper specification limit is defined as:

$$Ztotal = \Phi^{-1}(p(\text{not conform})total)$$

- p(not conform) below: The probability of producing a defect product below the lower specification limit is defined as:

$$p(\text{not conform})below = \Phi(Zbelow)$$

- p(not conform) above: The probability of producing a defect product above the upper specification limit is defined as:

$$p(\text{not conform})above = \Phi(Zabove)$$

- p(not conform) total: The probability of producing a defect product below or above the specification limits is defined as:

$$p(\text{not conform})total = p(\text{not conform})below + p(\text{not conform})above$$

- PPM below: The number of defective products below the lower specification limit over one million items produced is defined as:

$$PPMbelow = p(notconform)below \times 10^6$$

- PPM above: The number of defective products above the upper specification limit over one million items produced is defined as:

$$PPMabove = p(notconform)above \times 10^6$$

- PPM total: The number of defective products below or above the specification limits over one million items produced is defined as:

$$PPMtotal = PPMbelow + PPMabove$$

Box-Cox transformation

Box-Cox transformation is used to improve the normality of the time series; the Box-Cox transformation is defined by the following equation:

$$Y_t = \begin{cases} \frac{X_t^\lambda - 1}{\lambda}, & (X_t \geq 0, \lambda \neq 0) \text{ ou } (X_t \geq 0, \lambda > 0), \\ ln(X_t), & X_t > 0, \lambda = 0, \end{cases}$$

Where the series X_t being transformed into series Y_t, $t = 1, \ldots, n$.

Note: if $\lambda < 0$, the first equation is still valid, but X_t must be strictly positive. XLSTAT accepts a fixed value of λ, or it can find the value that maximizes the likelihood value, the model being a simple linear model with the time as sole explanatory variable.

Chart rules

XLSTAT lets you apply rules for special causes and Westgard rules on the data set. Two sets of rules are available in order to interpret control charts. You can activate and deactivate the rules separately in each set.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. Below are descriptions of the various elements of the dialog box.

OK: Click this button to start the calculations.

Cancel: Click this button to close the dialog box without doing any calculations.

: Click this button to display help options.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Chart type: Select the type of chart you want to use (see the description section for more details):

- P chart
- NP chart
- C chart
- U chart

Data: Please choose the unique column or row that contains all the data.

Groups: Activate this Option to select a column/row that contains the group identifier. Select the data that identify the corresponding group for each element of the data selection .

Common subgroup size: If the subgroup size is constant, then you can deactivate the groups option and enter the common subgroup size in this field.

Phase: Activate this option to supply one column/row with the phase identifier.

- **Different specifications**: Activate this option if you want to enter specifications specific to each phase for the process capability parameters. In this case, in the Options tab, you must enter a USL value, an LSL value and a target value for each phase.

Observation labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Variable labels" option is activated, you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet in the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row (column mode) or column (row mode) of the data selections contains a label.

Options tab:

Upper control limit (UCL):

Bound: Activate this option, if you want to enter a maximum value to accept for the upper control limit of the process. This value will be used when the calculated upper control limit is greater than the value entered here.

Value: Enter the upper control limit. This value will be used and overrides the calculated upper control limit.

Lower control limit (LCL):

Bound: Activate this option, if you want to enter a minimum value to accept for the lower control limit of the process. This value will be used when the calculated lower control limit is greater than the value entered here.

Value: Enter the lower control limit. This value will be used and overrides the calculated upper control limit.

Calculate Process capabilities: Activate this option to calculate process capabilities based on the input data (see the description section for more details).

USL: If the calculation of the process capabilities is activated, please enter the upper specification limit (USL) of the process here.

LSL: If the calculation of the process capabilities is activated, please enter the lower specification limit (LSL) of the process here.

Target: If the calculation of the process capabilities is activated, activate this option to add the target value of the process.

Confidence interval (%):If the "Calculate Process Capabilities" option is activated, please enter the percentage range of the confidence interval to use for calculating the confidence interval around the parameters. Default value: 95.

Box-Cox transformation: Activate this option to compute the Box-Cox transformation. You can either fix the value of the Lambda parameter, or decide to let *XL* STAT optimize it (see the description for further details).

k Sigma: Activate this option to enter the distance between the upper and the lower control limit and the center line of the control chart. The distance is fixed to k times the factor you enter, multiplied by the estimated standard deviation. Corrective factors according to Burr (1969) will be applied.

alpha (%): Activate this option to enter the size of the confidence range around the center line of the control chart. The alpha is used to compute the upper and lower control limits. 100 – alpha % of the distribution of the control chart is inside the control limits. Corrective factors according to Burr (1969) will be applied.

P bar / C bar / U bar: Activate this option to enter a value for the center line of the control chart. This value should be based on historical data.

Outputs tab:

Display zones: Activate this option to display the limits of the zones A and B beside the lower and upper control limit as well.

Normality Tests: Activate this option to check normality of the data. (see the Normality Tests tool for further details).

Significance level (%): Enter the significance level for the tests.

Test special causes: Activate this option to analyze the points of the control chart according to the rules for special causes. You can activate the following rules independently:

- 1 point more than 3s from center line.
- 9 points in a row on same side of center line.
- 6 points in a row, all increasing or all decreasing.
- 14 points in a row, alternating up and down.
- 2 out of 3 points > 2s from center line (same side).
- 4 out of 5 points > 1s from center line (same side).
- 15 points in a row within 1s of center line (either side).
- 8 points in a row > 1s from center line (either side).
- **All**: Click this button to select all.
- **None**: Click this button to deselect all.

Apply Westgard rules: Activate this option to analyze the points of the control chart according to the Westgard rules. You can activate the following rules independently:

- Rule 1 2s.
- Rule 1 3.
- Rule 2 2s.
- Rule 4s.
- Rule 4 1s.
- Rule 10 X.
- **All**: Click this button to select all.
- **None**: Click this button to deselect all.

Charts tab:

Display charts: Activate this option to display the control charts graphically.

Normal Q-Q Charts: Check this option to display Q-Q plots.

Histograms: Activate this option to display the histograms of the samples. For a theoretical distribution, the density function is displayed.

Run Charts: Activate this option to display a chart of the latest data points. Each individual measurement is displayed.

Results

Estimation:

Estimated mean: This table displays the estimated mean values for the different phases.

Estimated standard deviation: This table displays the estimated standard deviation values for the different phases.

Box-Cox transformation:

Estimates of the parameters of the model: This table is available only if the Lambda parameter has been optimized. It displays the estimator for Lambda.

Series before and after transformation: This table displays the series before and after transformation. If Lambda has been optimized, the transformed series corresponds to the residuals of the model. If it hasn't, then the transformed series is the direct application of the Box-Cox transformation.

Process capabilities: These tables are displayed if the "process capability" option has been selected. There is one table for each phase. A table contains the following indicators for the process capability and, if possible, the corresponding confidence intervals: Cp, Cpl, Cpu, Cpk, Pp, Ppl, Ppu, Ppk, Cpm, Cpm (Boyle), Cp 5.5, Cpk 5.5, Cpmk, and Cs (Wright).

For Cp, Cpl, and Cpu, information about the process performance is supplied and for Cp a status information is given to facilitate the interpretation.

Cp values have the following status based on Ekvall and Juran (1974):

- "not adequate" if $Cp < 1$,
- "adequate" if $1 \leq Cp \leq 1.33$,
- "more than adequate" if $Cp > 1.33$.

Based on Montgomery (2001), Cp needs to have the following minimal values for the process performance to be as expected:

- 1.33 for existing processes,
- 1.50 for new processes or for existing processes when the variable is critical,
- 1.67 for new processes when the variable is critical.

Based on Montgomery (2001), Cpu and Cpl need to have the following minimal values for process performance to be as expected:

- 1.25 for existing processes,
- 1.45 for new processes or for existing processes when the variable is critical.
- 1.60 for new processes when the variable is critical.

Capabilities: This chart contains information about the specification and control limits. A line between the lower und upper limits represents the interval with an additional vertical mark for the center line. The different control limits of each phase are drawn separately.

Chart information:

The following results are displayed separately for each requested chart. Charts can be selected alone or in combination with the X attribute chart.

P / NP / C / U chart: This table contains information about the center line and the upper and lower control limits of the selected chart. There will be one column for each phase.

Observation details: This table displays detailed information for each observation. For each observation the corresponding phase, the value for P, NP, C or U, the subgroup size, the center line, the lower and upper control limits are displayed. If the information about the zones A, B and C are activated, then the lower and upper control limits of the zones A and B are displayed as well.

Rule details: If the rules options are activated, a detailed table about the rules will be displayed. For each subgroup there is one row for each rule that applies. "Yes" indicates that the corresponding rule was fired, and "No" indicates that the rule does not apply.

P / NP / C / U Chart: If the charts are activated, then a chart containing the information of the two tables above is displayed. The center line and the lower and upper control limits are displayed as well. If the corresponding options have been activated, the lower and upper control limits for the zones A and B are included and there are labels for the subgroups for which rules were fired. A legend with the activated rules and the corresponding rule number is displayed below the chart.

Normality tests:

For each of the four tests, the statistics relating to the test are displayed including, in particular, the p-value which is afterwards used in interpreting the test by comparing with the chosen significance threshold.

If requested, a Q-Q plot is then displayed.

Histograms: The histograms are displayed. If desired, you can change the color of the lines, scales, titles as with any Excel chart.

Run chart: The chart of the last data points is displayed.

Example

A tutorial explaining how to use the attributes charts tool is available on the XLSTAT Help Center. To consult the tutorial, please go to:

http://www.xlstat.com/demoSPA_EN.htm

References

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.

Time Weighted Charts

Use this tool to supervise production quality where you have a group of measurements or a single measurement for each point in time. The measurements need to be quantitative variables.

This tool is useful to recap the mean and the variability of the measured production quality.

This tool includes integrated Box-Cox transformations, calculation of process capability and the application of rules for special causes and Westgard rules (an alternative rule set to identify special causes) available to complete your analysis.

In this section:

Description

Control charts were first mentioned in a document Walter Shewhart wrote during his time working at Bell Labs in 1924. He described his methods completely in his book (1931).

For a long time, there was no significant innovation in the area of control charts. With the development of CUSUM, UWMA and EWMA charts in 1936, Deming expanded the set of available control charts.

Control charts were originally used in the factory production of goods. Therefore, the wording is still from that domain. Today this approach is being applied to a large number of different fields, including services, human resources, and sales. In the following chapters, we will use the wording from the production and shop floors.

Time Weighted charts

The time weighted charts tool offers you the following chart types:

- CUSUM or CUSUM individual
- UWMA or UWMA individual
- EWMA or EWMA individual

A CUSUM, UWMA or EWMA chart lets you follow the mean of a production process. Mean shifts are easily visible in the diagrams.

UWMA and EWMA charts

These charts are not directly based on the raw data. They are based on the smoothed data.

In the case of UWMA charts, data are smoothed using a uniform weighting in a moving window. Then the chart is analyzed like Shewhart charts.

In the case of EWMA charts, the data is smoothed using exponentially decreasing weighting. Then the chart is analyzed like Shewhart charts.

CUSUM charts

These charts are not directly based on the raw data. They are based on the normalized data.

These charts help to detect mean shifts of a granularity defined by the user. The granularity is defined by the design parameter k. k is the half of the mean shift to be detected. To detect a 1 sigma shift, k is set to 0.5.

Two kinds of CUSUM charts can be drawn: one and two sided charts. In the case of a one sided CUSUM chart, upper and lower cumulated sums SH and SL are recursively calculated.

$$SH_i = max(0, (z_i - k) + SH_{i-1})$$

$$SL_i = min(0, (z_i + k) + SL_{i-1})$$

If SH or SL is bigger than the threshold h, then a mean shift is detected. The value of h can be chosen by the user (h is usually set to 4 or 5).

The initial value of SH and SL at the beginning of the calculation and after detecting a mean shift is usually 0. Using the option FIR (Fast Initial Response) can change this initial value to a user defined value.

In the case of a two sided CUSUM chart the normalized data are calculated. The upper and lower control limits are called "U mask" or "V mask". These names are related to the shape that the control limits draws on the chart. For a given data point, the maximal upper and lower limits for mean shift detection are calculated backwards and drawn in the chart in a U or V mask format. The default data point for the origin of the mask is the last data point. The user can change this by the option origin.

This tool offers you the following options for the estimation of the standard deviation (sigma) of the data set, given k subgroups and n_i $(i = 1, \ldots, k)$ measurements per subgroup:

- **Pooled standard deviation**: sigma is computed using the k within-subgroup variances:

$$\hat{s} = \sqrt{\frac{\sum_{i=1}^{k} (n_i - 1)\, s_i^2}{\sum_{i=1}^{k} (n_i - 1)}} / c_4 \left(1 + \sum_{i=1}^{k} (n_i - 1)\right)$$

- **R bar**: The estimator for sigma is calculated based on the average range of the k subgroups. $\hat{s} = \overline{R}/d_2$ where d_2 is the control chart constant according to Burr (1969).
- **S bar**: The estimator for sigma is calculated based on the average of the standard deviations of the k subgroups: $\hat{S} = \sqrt{\frac{1}{k}\sum_{i=1}^{k} s_i^2}/c_4$

In the case of n Individual measurements:

- **Average moving range**: The estimator for sigma is calculated based on the average moving range using a window length of m measurements. $\hat{s} = \overline{m}/d_2$, where d_2 is the control chart constant according to Burr (1969).
- **Median moving range**: The estimator for sigma is calculated based on the median of the moving range using a window length of m measurements. $\hat{s} = \overline{median}/d_4$ where d_4 is the control chart constant according to Burr (1969).
- **standard deviation**: The estimator for sigma is calculated based on the standard deviation of the n measurements: $\hat{s} = s/c_4$ where c_4 is the control chart constant according to Burr (1969).

Process capability

Process capability describes a process and determines whether the process is under control and the distributions of the measured variables are inside the specification limits of the process. If the distributions of the measured variables are in the technical specification limits, then the process is called "capable".

During the interpretation of the different indicators for the process capability, please pay attention to the fact that some indicators suppose normality or at least symmetry of the distribution of the measured values. By using a normality test, you can verify these premises (see the Normality Tests in XLSTAT-Pro).

If the data are not normally distributed, you have the following possibilities to obtain results for the process capabilities:

- Use the Box-Cox transformation to improve the normality of the data set. Then verify the normality again using a normality test.
- Use the process capability indicator Cp 5.5.

Let m, s, LSL, USL be (respectively) the estimated mean, standard deviation, lower specification limit, upper specification limit of the process, and let T be the selected target. Let μ and σ be the true mean and standard deviation of the process. XLSTAT allows to compute the following performance indicators to evaluate the process capability:

- Cp: The short term process capability is defined as:

$$Cp = \frac{USL - LSL}{6s}$$

- Cpl: The short term process capability with respect to the lower specification is defined as:

$$Cpl = \frac{m - LSL}{3s}$$

- Cpu: The short term process capability with respect to the upper specification is defined as:

$$Cpu = \frac{USL - m}{3s}$$

- Cpk: The short term process capability supposing a centered distribution is defined as:

$$Cpk = min(Cpl, Cpu)$$

- Pp: The long term process capability is defined as:

$$Pp = \frac{USL - LSL}{6\sigma}$$

- Ppl: The long term process capability with respect to the lower specification is defined as:

$$Ppl = \frac{m - LSL}{3\sigma}$$

- Ppu: The long term process capability with respect to the upper specification is defined as:

$$Ppu = \frac{USL - m}{3\sigma}$$

- Ppk: The long term process capability supposing a centered distribution is defined as:

$$Ppk = min(Ppl, Ppu)$$

- Cpm: The short term process capability according to Taguchi. This value can be calculated, if the target value has been specified. It is defined as:

$$Cpm = \frac{USL - LSL}{6\sqrt{s^2 + (m - T)^2}}$$

- Cpm Boyles: The short term process capability according to Taguchi improved by Boyles (1991). This value can be calculated if the target value has been specified. It is defined as:

$$Cpm\ Boyles = \frac{USL - LSL}{6\sqrt{(n-1)s^2/n + (m - T)^2}}$$

- Cp 5.15: The short term process capability is defined as:

$$Cp\ 5.15 = \frac{USL - LSL}{5.15s}$$

- Cpk 5.15: The short term process capability supposing a centered distribution is defined as:

$$Cpk\ 5.15 = \frac{(USL - LSL)/2 - |m - (USL + LSL)/2|}{2.57s}$$

- Cpmk: The short term process capability according to Pearn. This value can be calculated, if the target value has been specified. It is defined as:

$$Cpmk = \frac{s \cdot Cpk}{\sqrt{s^2 + (m - T)^2}}$$

- Cs Wright: The process capability according to Wright. This value can be calculated, if the target value has been specified. It is defined as:

$$Cs = \frac{(USL - LSL)/2 - |m - T|}{3\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - T)^2 + \left|\frac{n^2 m_3}{(n-1)(n-2)} \times \left(\frac{n}{n-1} \times \frac{m_2}{c_4^2}\right)^{-1/2}\right|}}$$

with

$$m_r = \sum_{i=1}^{n}(x_i - T)^r$$

$$c_4 = \left[\frac{2}{n-1}\right]^{1/2} \Gamma(\frac{n}{2})/\Gamma(\frac{n-1}{2})^{-1}$$

- Z below: The amount of standard deviations between the mean and the lower specification limit is defined as:

$$Zbelow = (m - LSL)/s$$

- Z above: The amount of standard deviations between the mean and the upper specification limit is defined as:

$$Zabove = (USL - m)/s$$

- Z total: The amount of standard deviations between the mean and the lower or upper respectively specification limit is defined as:

$$Ztotal = \Phi^{-1}(p(\text{not conform})total)$$

- p(not conform) below: The probability of producing a defective product below the lower specification limit is defined as:

$$p(\text{not conform})below = \Phi(Zbelow)$$

- p(not konform) above: The probability of producing a defective product above the upper specification limit is defined as:

$$p(\text{not conform})above = \Phi(Zabove)$$

- p(not conform) total: The probability of producing a defective product below or above the specification limits is defined as:

$$p(\text{not conform})total = p(\text{not conform})below + p(\text{not conform})above$$

- PPM below: The number of defective products below the lower specification limit over one million items produced is defined as:

$$PPMbelow = p(notconform)below \times 10^6$$

- PPM above: The number of defective products above the upper specification limit over one million items produced is defined as:

$$PPMabove = p(notconform)above \times 10^6$$

- PPM total: The number of defective products below or above the specification limits over one million items produced is defined as:

$$PPMtotal = PPMbelow + PPMabove$$

Box-Cox transformation: Box-Cox transformation is used to improve the normality of the time series; the Box-Cox transformation is defined by the following equation:

$$Y_t = \begin{cases} \frac{X_t^\lambda - 1}{\lambda}, & (X_t > 0, \lambda \neq 0) \text{ or } (X_t \geq 0, \lambda > 0) \\ \ln(X_t), & X_t > 0, \quad \lambda = 0 \end{cases}$$

Where the series $\{X_t\}$ being transformed into series $\{Y_t\}$, $(t = 1, \ldots, n)$:

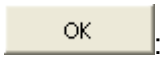
Note: if $l < 0$ the first equation is still valid, but X_t must be strictly positive. XLSTAT accepts a fixed value of l, or it can find the value that maximizes the likelihood value, the model being a simple linear model with the time as sole explanatory variable.

Chart rules

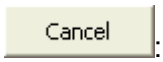
XLSTAT lets you apply rules for special causes and Westgard rules on the data set. Two sets of rules are available in order to interpret control charts. You can activate and deactivate the rules in each set separately.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. Below are descriptions of the various elements of the dialog box.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help options.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange sheet of paper, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Chart type: Select the type of chart you want to use (see the description section for more details):

- CUSUM chart
- CUSUM individual chart
- UWMA chart
- UWMA individual chart
- EWMA chart
- EWMA individual chart

Data format: Select the data format.

- **Columns/Rows**: Activate this option for XLSTAT to take each column (in column mode) or each row (in row mode) as a separate measurement that belongs to the same subgroup.
- **One column/Row**: Activate this option if the measurements of subgroups continuously follow one after the other in one column or one row. To assign the different measurements to their corresponding subgroup, please enter a constant group size or select a column or row with the group identifier in it.

Data: If the data format “One column/row” is selected, please choose the unique column or row that contains all the data. The assignment of the data to their corresponding subgroup must be specified using the Groups field or by setting the common subgroup size. If you select the data « Columns/rows » option, please select a data area with one column/row per measurement in a subgroup.

Groups: If the data format « one column/row » is selected, then activate this Option to select a column/row that contains the group identifier. Select the data that identifies for each element of the data selection the corresponding group.

Common subgroup size: If the data format “One column/row” is selected and the subgroup size is constant, then you can deactivate the groups option and enter in this field the common subgroup size.

Phase: Activate this option to supply one column/row with the phase identifier.

- **Different specifications**: Activate this option if you want to enter specifications specific to each phase for the process capability parameters. In this case, in the Options tab you must enter a USL value, an LSL value and a target value for each phase.

Observation labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Variable labels" option is activated, you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet in the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row (column mode) or column (row mode) of the data selections contains a label.

Standardize: In the case of a CUSUM chart, please activate this option to display the cumulated sums and the control limits normalized.

Target: In the case of a CUSUM chart, please activate this option to enter the target value that will be used during the normalization of the data. The default value is the estimated mean.

Weight: In the case of a EWMA chart, please activate this option to enter the weight factor of the exponential smoothing.

MA Length: In the case of a UWMA chart, please activate this option to enter the length of the window of the moving average.

Options tab:

Upper control limit:

Bound: Activate this option, if you want to enter a maximum value to accept for the upper control limit of the process. This value will be used when the calculated upper control limit is greater than the value entered here.

Value: Enter the upper control limit. This value will be used and overrides the calculated upper control limit.

Lower control limit:

Bound: Activate this option, if you want to enter a minimum value to accept for the lower control limit of the process. This value will be used when the calculated lower control limit is greater than the value entered here.

Value: Enter the lower control limit. This value will be used and overrides the calculated upper control limit.

Calculate Process capabilities: Activate this option to calculate process capabilities based on the input data (see the description section for more details).

USL: If the calculation of the process capabilities is activated, please enter the upper specification limit (USL) of the process here.

LSL: If the calculation of the process capabilities is activated, please enter the lower specification limit (LSL) of the process here.

Target: If the calculation of the process capabilities is activated, activate this option to add the target value of the process.

Confidence interval (%): If the "Calculate Process Capabilities" option is activated, please enter the percentage range of the confidence interval to use for calculating the confidence interval around these parameters. Default value: 95.

Box-Cox: Activate this option to compute the Box-Cox transformation. You can either fix the value of the λ parameter, or decide to let XLSTAT optimize it (see the description for further details).

k Sigma: Activate this option to enter the distance between the upper and the lower control limit and the center line of the control chart. The distance is fixed to k times the factor you enter multiplied by the estimated standard deviation. Corrective factors according to Burr (1969) will be applied.

alpha: Activate this option to enter the size of the confidence range around the center line of the control chart. The alpha is used to compute the upper and lower control limits. $(100 - \alpha)\%$ of the distribution of the control chart is inside the control limits. Corrective factors according to Burr (1969) will be applied.

Mean: Activate this option to enter a value for the center line of the control chart. This value should be based on historical data.

Sigma: Activate this option to enter a value for the standard deviation of the control chart. This value should be based on historical data. If this option is activated, then you cannot choose an estimation method for the standard deviation in the "Estimation" tab.

Estimation tab:

Method for Sigma: Select an option to determine the estimation method for the standard deviation of the control chart (see the description for further details):

- Pooled standard deviation: The standard deviation is calculated using all available measurements. That means having n subgroups with k measurements for each subgroup, all the n * k measurements will be weighted equally to calculate the standard deviation.
- R bar: The estimator of sigma is calculated using the average range of the n subgroups.
- S bar: The estimator of sigma is calculated using the average standard deviation of the n subgroups.
- Average Moving Range: The estimator of sigma is calculated using the average moving range using a window length of m measurements.
- Median Moving Range: The estimator of sigma is calculated using the median of the moving range using a window length of m measurements.
- MR Length: Activate this option to change the window length of the moving range.
- Standard deviation: The estimator of sigma is calculated using the standard deviation of the n measurements.

Design tab:

This tab is only active if CUSUM charts are selected.

Scheme: Choose one of the following options depending on the kind of chart that you want (see the description for further details):

- **One sided (LCL/UCL)**: The upper and lower cumulated sum are calculated separately for each point.
- **FIR**: Activate this option to change the initial value of the upper and lower cumulated sum. Default value is 0.
- **Two sided (U-Mask)**: The normalized values are displayed. Starting from the origin point, the upper and lower limits for the mean shift detection are displayed backwards in the form of a mask.
- **Origin**: Activate this option to change the origin of the mask. Default value is the last data point.

Design: In this section you can determine the Parameter of the mean-shift detection (see the description for further details):

- **h**: Enter the threshold for the upper and lower cumulated sum or the mask from above which a mean shift is detected.
- **k**: Enter the granularity of the mean shift detection. K is the half of the mean shift to be detected. Default value is 0.5 to detect 1 sigma mean shifts.

Outputs tab:

Display zones: Activate this option to display the limits of the zones A and B beside the lower and upper control limit.

Normality Tests: Activate this option to check normality of the data. (see the Normality Tests tool for further details).

Significance level (%): Enter the significance level for the tests.

Test special causes: Activate this option to analyze the points of the control chart according to the rules for special causes. You can activate the following rules independently:

- 1 point more than 3s from center line
- 9 points in a row on same side of center line
- 6 points in a row, all increasing or all decreasing
- 14 points in a row, alternating up and down
- 2 out of 3 points > 2s from center line (same side)
- 4 out of 5 points > 1s from center line (same side)
- 15 points in a row within 1s of center line (either side)
- 8 points in a row > 1s from center line (either side)
- **All**: Click this button to select all.
- **None**: Click this button to deselect all.

Apply Westgard rules: Activate this option to analyze the points of the control chart according to the Westgard rules. You can activate the following rules independently:

- Rule 1 2s
- Rule 1 3
- Rule 2 2s
- Rule 4s
- Rule 4 1s

- Rule 10 X
- **All**: Click this button to select all.
- **None**: Click this button to deselect all.

Charts tab:

Display charts: Activate this option to display the control charts graphically.

Normal Q-Q Charts: Check this option to display Q-Q plots.

Histograms: Activate this option to display the histograms of the samples. For a theoretical distribution, the density function is displayed.

Run Charts: Activate this option to display a chart of the latest data points. Each individual measurement is displayed.

Results

Estimation:

Estimated mean: This table displays the estimated mean values for the different phases.

Estimated standard deviation: This table displays the estimated standard deviation values for the different phases.

Box-Cox transformation:

Estimates of the parameters of the model: This table is available only if the Lambda parameter has been optimized. It displays the estimator for Lambda.

Series before and after transformation: This table displays the series before and after transformation. If Lambda has been optimized, the transformed series corresponds to the residuals of the model. If it hasn't then the transformed series is the direct application of the Box-Cox transformation

Process capabilities:

Process capabilities: These tables are displayed, if the "process capability" option has been selected. There is one table for each phase. A table contains the following indicators for the process capability and if possible the corresponding confidence intervals: Cp, Cpl, Cpu, Cpk, Pp, Ppl, Ppu, Ppk, Cpm, Cpm (Boyle), Cp 5.5, Cpk 5.5, Cpmk, and Cs (Wright).

For Cp, Cpl, and Cpu, information about the process performance is supplied and for Cp status information is given to facilitate the interpretation.

Cp values have the following status based on Ekvall and Juran (1974):

- "not adequate" if $Cp < 1$

- "adequate" if $1 \leq Cp \leq 1.33$
- "more than adequate" if $Cp > 1.33$

Based on Montgomery (2001), Cp needs to have the following minimal values for the process performance to be as expected:

- 1.33 for existing processes
- 1.50 for new processes or for existing processes when the variable is critical
- 1.67 for new processes when the variable is critical

Based on Montgomery (2001), Cpu and Cpl need to have the following minimal values for process performance to be as expected:

- 1.25 for existing processes
- 1.45 for new processes or for existing processes when the variable is critical
- 1.60 for new processes when the variable is critical

Capabilities: This chart contains information about the specification and control limits. A line between the lower und upper limits represents the interval with an additional vertical mark for the center line. The different control limits of each phase are drawn separately.

Chart information:

The following results are displayed separately for the requested chart.

UWMA / EWMA / CUSUM chart: This table contains information about the center line and the upper and lower control limits of the selected chart. There will be one column for each phase.

Observation details: This table displays detailed information for each subgroup. For each subgroup in the corresponding phase, the values according to the selected diagram type, the center line, the lower and upper control limits are displayed. If the information about the zones A, B and C are activated, then the lower and upper control limits of the zones A and B are displayed as well.

Rule details: If the rules options are activated, a detailed table about the rules will be displayed. For each subgroup, there is one row for each rule that applies. "Yes" indicates that the corresponding rule was fired, and "No" indicates that the rule does not apply.

UWMA / EWMA / CUSUM Chart: If the charts are activated, then a chart containing the information from the two tables above is displayed. The center line and the lower and upper control limits are displayed as well. If the corresponding options have been activated, the lower and upper control limits for the zones A and B are included and there are labels for the subgroups for which rules were fired. A legend with the activated rules and the corresponding rule number is displayed below the chart.

Normality tests:

For each of the four tests, the statistics relating to the test are displayed, including, in particular, the p-value which is later used in interpreting the test by comparing it with the chosen significance threshold.

If requested, a Q-Q plot is then displayed.

Histograms: The histograms are displayed. If desired, you can change the color of the lines, scales, titles as with any Excel chart.

Run chart: The chart of the last data points is displayed.

Example

A tutorial explaining how to use the SPC time weighted charts tool is available on the XLSTAT Help Center. To consult the tutorial, please go to:

http://www.xlstat.com/demoSPW_EN.htm

References

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.

Pareto plots

Use this tool to calculate descriptive statistics and display Pareto plots for a set of causes (or qualitative variables).

In this section:

Description

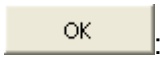
A Pareto chart draws its name from an Italian economist, but J. M. Juran is credited with being the first to apply it to industrial problems.

The causes that should be investigated (e. g., nonconforming items) are listed and percentages assigned to each one so that the total is 100 %. The percentages are then used to construct the diagram that is essentially a sorted bar chart and the associated cumulative curve. Pareto analysis uses the ranking causes to determine which of them should be pursued first.

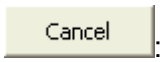
Although you can select several variables (or samples) at the same time, XLSTAT calculates all the descriptive statistics for each of the samples independently. A chart combining the different sub-samples may be displayed.

Dialog box

The dialog box is made up of several tabs corresponding to the various options for controlling the calculations and displaying the results. A description of the various components of the dialog box are given below.



: Click this button to start the calculations.



: Click this button to close the dialog box without doing any calculations.

: Click this button to display help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Causes: Select a column (or a row in row mode) of qualitative data that represent the list of causes you want to calculate descriptive statistics for.

Frequencies: Check this option, if your data is already aggregated in a list of causes and a corresponding list of frequencies of these causes. Select here the list of frequencies that correspond to the selected list of causes.

Sub-samples: Check this option to select a column showing the names or indexes of the sub-samples for each of the observations. * **Variable-Category labels**: Check this option if you want to display the subsamples specifying the variable name followed by the category name. * **Compare to total sample**: Check this option if you want to compare the sub-samples to the full sample.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selections contains a label.

Weights: Check this option if the observations are weighted. If you do not check this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Sample labels" option is activated.

Options tab:

Sorting options: Choose the sorting option among:

- **No sorting**: The causes are not sorted.
- **Descending**: The causes are sorted in descending order according to their frequencies.
- **First descending**: The causes are sorted in descending order according to the first series of frequencies.
- **Alphabetical**: The causes are sorted in ascending alphabetical order.

Combine categories: Activate this option if you want to combine some causes.

- **Frequency less than**: Choose this option to combine categories having a frequency smaller that the user defined value.
- **% smaller than**: Choose this option to combine categories having a % smaller that the user defined value.
- **K smallest categories**: Choose this option to combine the k smallest categories. The value k is defined by the user.
- **Cumulative %**: Choose this option to combine all categories, as soon as the cumulative % of the Pareto plot is bigger than the user defined value.

Charts tab:

Values used: Choose the type of data to be displayed on the left ordinates axis:

- **Frequencies**: Choose this option to make the scale of the plots correspond to the frequencies of the categories.
- **Relative frequencies**: Choose this option to make the scale of the plots correspond to the relative frequencies of the categories.

Change color at: Activate this option if you want to change the color of the bars when a specific cumulative % is reached.

All series on one chart: If you selected multiples series of causes or multiple frequencies, activate this option to display all series on a single summary chart.

Example

An example showing how to create Pareto charts is available on XLSTAT Help Center:

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

References

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 for quantitative variables (Measurement System Analysis)

Use this tool to control and validate your measurement method and measurement systems, in cases where you have several quantitative measures taken by one or more operators on several parts.

In this section:

Description

Measurement System Analysis (MSA) or Gage R&R (Gage Repeatability and Reproducibility) is a method to control and judge a measurement process. It is useful to determine which sources are responsible for variation in the measurement data. Variability can be caused by the measurement system, the operator or the parts. Gage R&R applied to quantitative measurements is based on two common methods: ANOVA and R control charts.

The word "gage" (or gauge) refers to the fact that the methodology is aimed at validating instruments or measurement methods.

A measurement is "repeatable" if the measures taken repeatedly by a given operator for the same object (product, unit, part, or sample, depending of the field of application) do not vary above a given threshold. If the repeatability of a measurement system is not satisfactory, one should question the quality of the measurement system, or train the operators that do not obtain repeatable results if the measurement system does not appear to be responsible for the high variability.

A measurement is "reproducible" if the measures obtained for a given object (product, unit, part, or sample, depending on the field of application) by several operators do not vary above a given threshold. If the reproducibility of a measurement system is not satisfactory, one should train the operators so that their results are more homogeneous.

The goal of a Gage R&R analysis is to identify the sources of variability and to take the necessary actions to reduce them if needed.

When the measures are quantitative data, two alternative methods are available for Gage R&R analysis. This is based on analysis of variance (ANOVA) and on R control charts (Range and average).

In the descriptions below, $\hat{\sigma}^2_{Repeatability}$ stands for the variance corresponding to repeatability. The lower it is, the more repeatable the measurement is (an operator gives coherent results for a given part). Its computation is different for the ANOVA and for the R control charts. $\hat{\sigma}^2_{\text{Reproducibility}}$ is the fraction of the total variance that corresponds to reproducibility. The lower it is, the more reproducible the measurement is (the various operators making consistent measurements for a given part). Its computation is different for the ANOVA and for the R control charts.

$\hat{\sigma}^2_{R\&R}$ is the variance of the gage R&R. The computation is always the sum of the two previous variances $\hat{\sigma}^2_{R\&R} = \hat{\sigma}^2_{Repeatability} + \hat{\sigma}^2_{\text{Reproducibility}}$.

ANOVA:

When the ANOVA model is used in R&R analysis, one can statistically test whether the variability of the measurements is related to the operators, and/or to the parts being measured themselves, and/or to an interaction between both (some operators might give significantly higher or lower measurements for some parts), or not. Two designs are available when doing gage R&R analysis: the crossed design (balanced) and the nested design.

Crossed design:

A balanced ANOVA with the two factors Operator and Part is carried out. You can choose between a reduced ANOVA model that involves only the main factors, or a full model that includes the interaction term as well (Part*Operator).

For a crossed ANOVA, the data must satisfy the needs of a balanced ANOVA. That means that for a given factor, you have equal frequencies for all categories, and each operator must have measured each part. In the case of a full ANOVA, the F statistics are calculated as follows:

$$F_{operator} = MSE_{operator} / MSE_{part*operator}$$

$$F_{part} = MSE_{part} / MSE_{part*operator}$$

where MSE stands for mean squared error.

If the p-Value of the interaction Operator*Part is bigger or equal to the user defined threshold (usually 25 %), the interaction term is removed from the model. We then have a reduced model.

In the case of a crossed ANOVA with interaction, the variances are defined as follows:

$$\hat{\sigma}^2 = MSE_{Error}$$
$$\hat{\sigma}^2_{part*operator} = \left(MSE_{part*operator} - MSE_{Error}\right)/nRep$$
$$\hat{\sigma}^2_{operator} = \left(MSE_{operator} - MSE_{part*operator}\right)/(nPart.nRep)$$
$$\hat{\sigma}^2_{part} = \left(MSE_{part} - MSE_{part*operator}\right)/(nOperator.nRep)$$
$$\hat{\sigma}^2_{Repeatability} = \hat{\sigma}^2$$
$$\hat{\sigma}^2_{\text{Reproducibility}} = \hat{\sigma}^2_{operator} + \hat{\sigma}^2_{part*operator}$$
$$\hat{\sigma}^2_{R\&R} = \hat{\sigma}^2_{Repeatability} + \hat{\sigma}^2_{\text{Reproducibility}}$$

In the case of a reduced model (without interaction), the variances are defined as follows:

$$\hat{\sigma}^2 = MSE_{Error}$$
$$\hat{\sigma}^2_{part*operator} = 0$$
$$\hat{\sigma}^2_{operator} = \left(MSE_{operator}\right)/(nPart.nRep)$$
$$\hat{\sigma}^2_{part} = \left(MSE_{part}\right)/(nOperator.nRep)$$
$$\hat{\sigma}^2_{Repeatability} = \hat{\sigma}^2$$
$$\hat{\sigma}^2_{\text{Reproducibility}} = \hat{\sigma}^2_{operator} + \hat{\sigma}^2_{part*operator}$$
$$\hat{\sigma}^2_{R\&R} = \hat{\sigma}^2_{Repeatability} + \hat{\sigma}^2_{\text{Reproducibility}}$$

where MSE stands for mean squared error, $nRep$ is the number of repetitions, $nPart$ is the number of parts, and $nOperator$ is the number of operators.

Nested design:

A nested ANOVA with the two factors Operator and Part(Operator) is carried out.

For a nested ANOVA, the data must satisfy the following prerequisites: for a given factor, you must have equal frequencies for all categories, and a part is checked by only one operator. The F statistics are calculated as follows:

$$F_{operator} = MSE_{operator}/MSE_{part(operator)}$$

$$F_{part(operator)} = MSE_{part(operator)}/MSE_{Error}$$

where MSE stands for mean squared error.

$$\hat{\sigma}^2 = MSE_{Error}$$
$$\hat{\sigma}^2_{Repeatability} = \hat{\sigma}^2$$
$$\hat{\sigma}^2_{\text{Reproducibility}} = \left(MSE_{operator} - MSE_{part(operator)}\right)/(nPart.nRep)$$
$$\hat{\sigma}^2_{R\&R} = \hat{\sigma}^2_{Repeatability} + \hat{\sigma}^2_{\text{Reproducibility}}$$

where MSE stands for mean squared error, $nRep$ is the number of repetitions, $nPart$ is the number of parts, and $nOperator$ is the number of operators.

R charts:

While less powerful than the ANOVA method, the Gage R&R analysis, based on Range and Average analysis, is easy to compute and produces control charts (R charts). Like the ANOVA method, it allows you to compute the repeatability and the reproducibility of the measurement process. To use this method, you need to have several parts, operators and repetitions (typically 10 parts, 3 operators, and 2 repetitions).

Based on the R chart, the different variances can be calculated as follows:

$$\hat{\sigma}^2_{Repeatability} = \bar{R}/d_2^*(nRep, nPart * nOperator)$$
$$\hat{\sigma}^2_{\text{Reproducibility}} = \left(\frac{\text{Max}(\mu_{Part}) - \text{Min}(\mu_{Part})}{d_2^*(nOperator, 1)}\right)^2 - \frac{\hat{\sigma}^2_{Repeatability}}{nPart * nOperator}$$
$$\hat{\sigma}^2_{R\&R} = \hat{\sigma}^2_{Repeatability} + \hat{\sigma}^2_{\text{Reproducibility}}$$
$$\hat{\sigma}^2_{part} = \left(\frac{\text{Max}(\mu_{Operator}) - \text{Min}(\mu_{Operator})}{d_2^*(nPart, 1)}\right)^2$$
$$\hat{\sigma}^2 = \hat{\sigma}^2_{R\&R} + \hat{\sigma}^2_{part}$$

where $\text{Max}(\mu_{Part} \text{ (respectively Operator)}) - \text{Min}(\mu_{Part} \text{ (respectively Operator)})$ is the difference between the maximum and the minimum across operators (respectively parts) of the averages for each part (respectively operators), $nRep$ is the number of repetitions, $nPart$ is the number of parts, $nOperator$ is the number of operators and $d_2^*(m, k)$ is the control chart constant according to Burr (1969).

During computation of the repeatability, we see that the mean amplitude of the Range chart is used. The variability of the parts and the reproducibility are based on the mean values of the $\bar{X}$ chart.

Indicators:

XLSTAT offers several indicators derived from the variances to describe the measurement system.

The study variation for the different sources is calculated as the product of the corresponding standard deviation of the source and the used defined factor k Sigma:

$$\text{Study variation} = k * \hat{\sigma}$$

The tolerance in percent is defined as the ratio of the variance in the study and the user defined tolerance:

$$\%\text{ tolerance} = \text{Study variation}/\text{tolerance}$$

The process sigma in percent is defined as ratio of the standard deviation of the source and the user defined historic process sigma:

$$\% \text{ process} = \text{standard deviation of the source}/\text{process sigma}$$

Precision to tolerance ratio (P/T):

$$P/T = \frac{k * \hat{\sigma}^2_{R\&R}}{tolerance}$$

Rho P (Rho Part):

$$\rho_{Part} = \frac{\hat{\sigma}^2_{Part}}{\hat{\sigma}^2}$$

Rho M:

$$\rho_M = \frac{\hat{\sigma}^2_{R\&R}}{\hat{\sigma}^2}$$

Signal to noise ratio (SNR):

$$SNR = \sqrt{\frac{2\rho_{Part}}{1-\rho_{Part}}}$$

Discrimination ratio (DR):

$$DR = \frac{1+\rho_{Part}}{1-\rho_{Part}}$$

Bias:

$$\text{Bias} = \mu_{Measurements} - \text{target}$$

Bias in percent:

$$\text{Bias } \% = (\mu_{Measurements} - \text{target})/\text{tolerance}$$

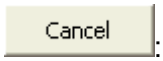
Resolution:

$$Resolution = Biais + 3 * \hat{\sigma}^2_{R\&R}$$

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files  .

General tab:

Y / Measurement: Choose the unique column or row that contains all the data. The assignment of the data to their corresponding subgroup must be specified using the Operator and the Parts field.

X / Operator: Select the data that identify for each element of the data selection the corresponding operator.

Parts: Select the data that identify for each element of the data selection the corresponding part.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row (column mode) or column (row mode) of the data selections contains a label.

Sort categories alphabetically: Activate this option to sort the categories of the variables in alphabetic order.

Options/Model tab:

Method: Choose the method to be used:

- **ANOVA**: Activate this option to calculate variances based on an ANOVA analysis.
- **R chart**: Activate this option to calculate variances based on an R chart.

k Sigma: Enter the user defined dispersion. Default value is 6.

Tolerance interval: Activate this option to define the amplitude of the tolerance interval (also $USL - LSL$).

Process sigma: Activate this option to enter a value for the standard deviation of the control chart. This value should be based on historical data.

Target: Activate this option to add the reference value of the measurements.

ANOVA: Choose the ANOVA model that should be used for the analysis:

- reduced
- crossed
- **Significance level (%)**: Enter the threshold below which the interaction of the crossed model should be taken into account. Default value is 5.
- nested

Options/Estimation tab:

Method for Sigma: Select the method for estimating the standard deviation of the control chart (see the description for further details):

- Pooled standard deviation
- R-bar
- S-bar

Outputs tab:

Variance components: Activate this option to show the table that displays the various variance components.

Status indicator: Activate this option to display the status indicators for the assessment of the measurement system.

Analysis of variance: Activate this option to display the variance analysis table.

Charts tab:

Display charts: Activate this option to display the control charts graphically.

- **Display zones**: Activate this option to display, beside the lower and upper control limit, the limits of the B and C zones.

Box plots: Check this option to display box plots (or box-and-whisker plots). See the description section of the univariate plots for more details.

Scattergrams: Check this option to display scattergrams. The mean (red +) and the median (red line) are always displayed.

- **Minimum/Maximum**: Check this option to systematically display the points corresponding to the minimum and maximum (box plots).
- **Outliers**: Check this option to display the points corresponding to outliers (box plots) with a hollowed-out circle.
- **Label position**: Select the position where the labels must be placed on the box plots and scattergrams plots.

Means charts: Activate this option to display the charts used to display the means of the various categories of the various factors.

Results

Variance components:

The first table and the corresponding chart display the variance split into its different sources. The contributions to the total variance and to the variance in the study, which is calculated using the user defined dispersion value, are given afterwards.

If a tolerance interval was defined, then the distribution of the variance by the variance according to the tolerance interval is displayed as well.

If a process sigma has been defined, then the distribution of the variance by the variance according to the process sigma is displayed as well.

The next table shows a detailed distribution of the variance by the different sources. Absolute values of the variance components and the percentage of the total variance are displayed.

The third table shows the distribution of the standard deviation for the different sources. It displays the absolute values of the variance components; the study variation that is calculated as the product of the standard deviation and the dispersion; the percentage of the study variation; the tolerance variability, which is defined as the ratio between variability of the study and the process sigma, and the percentage of the process variability.

Status indicator:

The first table shows information for the assessment of the measurement system. The Precision to tolerance ratio (P/T), Rho P, Rho M, Signal to noise ratio (SNR), Discrimination ratio (DR), absolute bias, and percentage and the resolution are displayed. The definition of the different indicators is given in the section description.

P/T values have the following status:

- "more than adequate" if P/T \<= 0.1
- "adequate" if 0.1 \< P/T \<= 0.3
- "not adequate" if P/T > 0.3

SNR values have the following status:

- "not acceptable" if SNR \< 2
- "not adequate" if 2 \<= SNR \<= 5
- "adequate" if SNR > 5

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.
- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).
- **R²**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2} \text{ where } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i$$

The R^2 is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer R^2 is to 1, the better the model is. The problem with the R^2 is that it does not consider the number of variables used to fit the model.

- **Adjusted R²**: The adjusted determination coefficient for the model. The adjusted R^2 can be negative if the R^2 is near zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}$$

The adjusted R^2 is a correction to the R^2 that considers the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W - p^*}\sum_{i=1}^{x} w_i(y_i - \hat{y}_i)^2$$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

Analysis of variance:

The **variance analysis table** is used to evaluate the explanatory power of the explanatory variables. The explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model whose independent variable would be a constant equal to the mean.

Chart information:

The following results are displayed separately for each requested chart. Charts can be selected alone or in combination with the X bar chart.

X bar/ R chart: This table contains information about the center line and the upper and lower control limits of the selected chart. There will be one column for each phase.

Observation details: This table displays detailed information for each subgroup (a subgroup corresponds to a pair of Operator*Part). For each subgroup the corresponding phase, the size, the mean, the minimum and the maximum values, the center line, and the lower and upper control limits are displayed. If the information about the zones A, B and C are activated, then the lower and upper control limits of the zones A and B are displayed as well.

X bar/ R chart: If the charts are activated, then a chart containing the information from the two tables above is displayed. Each subgroup is displayed. The center line and the lower and upper control limits are displayed as well. If the corresponding options have been activated, the lower and upper control limits for zones A and B are included, and there are labels for the subgroups for which rules were fired. A legend with the activated rules and the corresponding rule number is displayed below the chart.

Finally, the mean charts for each operator, for each part and for the interaction Operator*Part are displayed.

Example

A tutorial explaining how to use the Gage R&R tool is available on the XLSTAT Help Center. To consult the tutorial, please go to:

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

References

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 for Attributes (Measurement System Analysis)

Use this tool to control and validate your measurement method and measurement systems, in the case where you have qualitative measurements (attributes) nominal, or ordinal measurements taken by one or more operators on several parts.

In this section:

Description

Measurement System Analysis (MSA) or Gage R&R (Gage Repeatability and Reproducibility) is a method to control and judge a measurement process. It is useful to determine which sources are responsible for the variation of the measurement data. The word "gage" (or gauge) refers to the fact that the methodology is aimed at validating instruments or measurement methods.

Attributes data are qualitative characteristics (or attributes) than can be categorized and counted. Data can be nominal — that is, have multiple levels without natural ordering — or ordinal, that is, have at least three levels and natural ordering.

In contrast to the Gage R&R for quantitative measurements, the analysis based on attributes gives information on the "agreement" and on the "correctness". The concepts of variance, repeatability and reproducibility are not relevant in this case.

A high "agreement" of the measures taken by a given operator repeatedly for the same object (product, unit, part, or sample, depending on the field of application) shows that the operator is consistent. If the agreement of a measurement system is low, one should question the quality of the measurement system or protocol, or train the operators that do not obtain a high agreement, if the measurement system does not appear to be responsible for the lack of agreement.

A high "correctness" of the measures taken by an operator for the same object (product, unit, part, or sample, depending on the field of application) in comparison to the given reference or standard value shows that the operator comes to correct results. If the correctness of a measurement system is low, one should train the operators so that their results are more correct.

The goal of a Gage R&R analysis for attributes is to identify the sources of low agreement and low correctness, and to take the necessary actions.

The Gage R&R analysis for attributes is based on the following statistics to evaluate the agreement and correctness:

- Agreement statistics

- Kappa coefficients
- Kendall coefficients

If possible, the following comparisons are performed:

- Within operator
- Between operators
- Each operator versus reference
- All Operators versus reference

The reference (or standard) corresponds to the measurements reported by an expert or a method that is considered as highly reliable.

Agreement statistics:

It is possible to calculate these statistics in all the sections. In each section, the number of parts inspected, the number of matched scores and the ratio of matched scores, together with percentage of matched scores and confidence intervals, are displayed.

In the within-operator section, XLSTAT computes for each operator the number of cases where he agrees with himself for a given part across all repetitions.

In the between-operator section, XLSTAT computes the number of cases where all operators agree for a given part across all repetitions.

In the "Operator vs. reference" section, XLSTAT gives the number of cases where an operator agrees with the reference across all repetitions.

In the "all operators vs. reference section", XLSTAT computes the number of cases where all operators agree with the reference across all repetitions.

Kappa coefficients:

Cohen’s and Fleiss’ Kappa are well suited for qualitative variables. These coefficients are calculated on contingency tables obtained from paired samples. The Fleiss’ kappa is a generalization of the Cohen’s kappa. The kappa coefficient varies between -1 and 1. The higher the value of kappa, the stronger the agreement/correctness.

In the case of within-operator analysis, Cohen’s kappa can only be computed if there are exactly two trials.

In the case of between-operators analysis, Cohen’s kappa can only be computed for two operators with single trials.

Kendall coefficients:

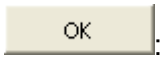
These indicators are available for ordinal variables with at least 3 categories.

The Kendall’s coefficient of concordance measures, on a 0 (no agreement) to 1 (perfect agreement) scale, the degree of concordance between two ordinal variables.

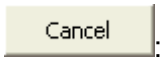
The Kendall's correlation coefficient, also referred to as Kendall tau or tau-b, allows you to measure on a -1 to 1 scale the degree of concordance between two ordinal variables.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options, ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help section.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select. If the button shows an orange paper sheet, XLSTAT displays additional buttons so that you can select data from flat files.  .

General tab:

Data format: Choose the data format:

- **Observations/variables table**: Activate this option if the data contain one column for all measurements.
- **Multiple columns**: Activate this option if the data contain one column per operator and repetition.

If the observations/variables table format is selected:

Measurement: Choose the unique column or row that contains all the measurement data.

Choose the data type:

- **Ordinal**: Activate this option if the measurement data is ordinal.
- **Nominal**: Activate this option if the measurement data is nominal.

Operator: Select the column or row that identifies the corresponding operator for each element of the data selection.

Parts: Select the column or row that identifies the corresponding part for each element of the data selection.

Reference: Activate this option if reference or standard values are available. Select the column or row that indicate the reference values for each measurement.

If the multiple column format is selected:

Measurement: Choose the table that contains all the measurement data.

Choose the data type:

- **Ordinal**: Activate this option if the measurement data is ordinal.
- **Nominal**: Activate this option if the measurement data is nominal.

Number of operator s: Enter the number of operators who evaluated the parts.

Number of repetitions: Enter the number of repetitions.

Reference: Activate this option if reference or standard values are available. Select the column or row that indicate for each measurement the reference values.

Operators name: activate this option if you know the name of each operator. Select the column or row that contains the names of each operator.

Range: Activate this option to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if the first row (column mode) or column (row mode) of the data selections contains a label.

Options tab:

Statistics:

- **Fleiss' kappa:** Activate this option if you want Fleiss' kappa to be computed
- **Cohen's kappa:** Activate this option if you want Cohen's kappa to be computed
- **Kendall's coefficient of concordance:** Activate this option if you want Kendall's coefficient of concordance to be computed (this option is only available for ordinal data)

- **Kendall’s correlation coefficient:** Activate this option if you want Kendall’s correlation coefficient to be computed (this option is only available for ordinal data and when a reference is available)
- **Confidence interval (%):** Enter the confidence interval (default value: 95)

Outputs tab:

Agreement: Activate this option to display the tables with the agreement statistics.

Intra operator: Activate this option to display the tables containing the intra operator results.

Between operators: Activate this option to display the tables containing the between operators' results.

Operator versus reference: Activate this option to display tables containing the operator vs. reference's results.

All operators versus reference: Activate this option to display tables containing the all operators vs. reference's results.

Charts tab:

Agreement charts: Activate this option to display the charts that show the mean values and their corresponding confidence intervals for the agreement statistics.

- **Within operator:** Activate this option to display the agreement chart for each operator over replicates.
- **Operator versus reference:** Activate this option to display the agreement chart for each operator compared to the reference.

Results

The tables with the selected statistics will be displayed.

The results are divided into the following four sections:

- Within operator
- Between operators
- Operator versus reference
- All Operators versus reference

Within each section, the following indicators are displayed, as far as the calculation is wanted and possible:

- Agreement statistics
- Kappa statistics
- Kendall statistics

References

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.

Design of Experiments

Screening designs

Use this module to generate a design to analyze the effect of 2 to 35 factors on one or more responses. This family of screening design is used to find the most influencing factors out of all the studied factors.

In this section:

Description

The family of screening designs aims for the study of the effect of two or more factors. In general, factorial designs are the most efficient for this type of study. But the number of necessary tests is often too large when using factorial designs. There are other possible types of designs in order to take into account the limited number of experiments that can be carried out.

XLSTAT proposes to carry out a full factorial design where all the combinations of factors will be carried out during the experiment. This type of design is very useful for having the best possible results, however, the number of experiments can very quickly increase. This is why XLSTAT integrates a large base of several hundred orthogonal design tables. Orthogonal design tables are preferred, as the ANOVA analysis will be based on a balanced design.

Based on the chosen number of factors and experiments, XLSTAT provides a secondary interface presenting a selection of orthogonal experimental designs closely aligned with the desired criteria (see Common designs). The orthogonal designs stored in XLSTAT and which can be proposed in the list include most of the classic designs: - full factorial designs - half factorial designs - quarter factorial designs - reduced factorial designs - Latin square designs - Plackett-Burman designs

In instances where the existing orthogonal designs fail to meet specific requirements, users have the option to explore D-Optimal designs. It is important to note, however, that these designs may not necessarily adhere to orthogonality principles.

At the output, in order to facilitate the entry of responses, individual sheets associated with each experience can be generated on separate Excel sheets, which can be printed and filled out.

Model

This tool generates designs that can be analyzed using an additive model without interactions for the estimation of the mean factor effects. If p is the number of factors, the ANOVA model is written as follows

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_{k(i,j),j} + \epsilon_i$$

Common designs

When starting the creation of an experimental design, the internal knowledge base is searched for common orthogonal designs that are close to the problem. A distance measure d between your problem and each common design is calculated in the following way:

p_i = number of factors with i categories in the problem

c_i = number of factors with i categories in the common design

p_{exp} = number of experiments in the problem

c_{exp} = number of experiments in the common design

$d(c,p) = \sum_{i=2}^{7} |c_i - p_i| + c_{exp} - p_{exp}$ (1)

All common designs having the same number of factors as the problem and having a distance $d < 20$ are proposed in a selection list.

The formal name for common designs is written in the following way:

$$L_n(p_1^{c_1}, ..., p_m^{c_m})$$

Where

n = number of experiments

c_i = number of categories of the group of factors p_i

p_i = number of factors having c_i categories

A common name for each design is displayed in the list if available.

D-Optimal design

To generate a D-Optimal design, XLSTAT uses Fedorov's algorithm, which uses a permutation method (see Cook and Nachtsheim, 1980). at each iteration, a simple exchange is carried out. The algorithm will then exchange the couple which optimizes the design according to the criterion described below.

The internal representation of the design matrix uses the following encoding. For a factor f_i having c_i categories, $c_i - 1$ columns $k_1, ..., k_{c_i - 1}$ are added in the design matrix X in the following way for the different category values of f_i:

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

The complete design matrix X is composed of n lines, where n is the number of experiences. The matrix contains a first column with 1 in each line and $c_i - 1$ columns for each factor f_i in the design, where c_i is the number of categories of the corresponding factor f_i.

X is the encoded design matrix, where every line represents the encoded experiment corresponding to the experimental design.

The criterion used for the optimization is defined as:

$$c = \log_{10}(\det(X^t X)) \tag{2}$$

With

$X^t X$ = information matrix

X = encoded design matrix

This criterion is named in the results as follows:

$$c = \mathrm{Log}(|I|)$$

The following common used criterion is also displayed in the results:

$$\mathrm{Log}(|I|^{1/p})$$

When comparing experimental designs that have a different number of experiences, the normalized log is used to be able to compare the different criteria values:

$$Norm.\log = \log_{10}((\det \frac{1}{N}(X^t X))^{1/p}) \tag{3}$$

This criterion is named in the results as follows:

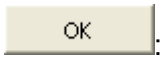
$$Norm.\log = \mathrm{Log}(|1/n * I|)^{1/p})$$

This measure allows comparing the optimality of different experimental designs, even if the number of experiences is different.

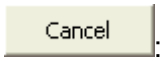
The user can select a number of repetitions in order to find a local optimum with good properties.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select.

General tab:

Quantitative factors (min/max): select the minimum and the maximum of the quantitative factors. The selection must have two rows, the first corresponds to the minimums, the second to the maximums for each of the factors.

Qualitative factors: select the qualitative factors and their modalities.

Number of responses: enter the number of responses to analyze.

Number of experiments: enter the number of experiments to be carried out during the design.

Repetitions: Activate this option to choose the number of repetitions of the design.

Randomize: Activate this option to change the order of the lines of the design into a random order.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections includes a header.

Options tab:

Method: Choose the method you want to use to generate the design.

- **Full factorial design**: select this option in order to create a full factorial design.
- **D-Optimal design**: This method allows to search for an optimal design.
- **Repetitions**: In the case of a random initial partition, enter the number of the repetitions to perform.
- **Stop conditions**:
  - **Iterations**: Enter the maximum number of iterations for the algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 50.
  - **Convergence**: Enter the maximum value of the evolution in the criterion from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.0001.
- **Orthogonal design**: this method makes it possible to find an orthogonal design close to the parameters entered by the user and present in the XLSTAT database.

Outputs tab:

Optimization summary: Activate this option to display the optimization summary.

Details of iterations: Activate this option to display the details of iterations.

Burt table: Activate this option to display the Burt table of the experimental design.

Encoded design: Activate this option to display the encoded experimental design in the case of a d-optimal design.

Field test: Activate this option to display the design in the form of plots.

Display experience sheets: Activate this option if you want to display individual Excel sheets for each experience. This can be useful in order to print them out and carry out the experiments.

Sort up: Activate this option to sort the categories in increasing order, the sort criterion being the value of the category. If this option is activated, the sort is ascending.

Sort the categories alphabetically: Activate this option so that the categories of all the variables are sorted alphabetically.

Variable-Category labels: Activate this option to use variable-category labels when displaying outputs. Variable-Category labels include the variable name as a prefix and the category name

as a suffix.

Charts tab:

3D view of the Burt table: Activate this option to display a 3D visualization of the Burt table.

Screening designs / Common designs dialog box:

Selection of experimental design: This dialog box lets you select the design of experiment you want to use. Thus, a list of fractional factorial designs is presented with their respective distance to the design that was to be generated. If you select a design and you click Select, then the selected design will appear. If no design fits your needs, click on the "optimize" button, and an algorithm will give you a design corresponding exactly to the selected factors.

Results

Variables information: This table shows the information about the factors. For each factor the short name, long name, unit and physical unit are displayed.

Experimental design: This table displays the complete experimental design. Additional columns include information on the factors and on the responses, a label for each experiment, the sort order, the run order and the repetition.

Responses optimization: The responses optimization table of is displayed after the experimental design. You must then select the following parameters:

- **Objective**: Choose the objective of the optimization. You have the choice between minimum, optimum and maximum.

If the selected objective is the optimum or the maximum, the following fields are activated:

- **Lower**: Enter for each answer the value of the lower bound below which the desirability is 0.
- **Target (left)**: Enter the value of the lower bound above which desirability is 1 for each response.

If the selected objective is the optimum or the minimum, the following fields are activated:

- **Target (right)**: Enter for each response the value of the upper bound below which the desirability is equal to 1.
- **Lower**: Enter for each answer the value of the upper limit above which the desirability is 0.
- **s**: Activate this option if the increasing desirability function must be non-linear. Then enter the value of the shape parameter which must be between 0.01 and 100.
- **t**: Activate this option if the decreasing desirability function must be non-linear. Then enter the value of the shape parameter which must be between 0.01 and 100.

- **Weight**: Activate this option if the answers must have an exponential value different from 1 when calculating desirability. Then enter the value of the shape parameter which must be between 0.01 and 100.

For more details on responses optimization, you can refer to the analysis of a factor effect design.

Endoded design: This table shows the encoded experimental design. This table is only displayed in the case of a d-optimal experimental design.

Burt table: The Burt table is displayed only if the corresponding option is activated in the dialog box. The **3D bar chart** that follows is the graphical visualization of this table.

If an Optimization was selected, then the following sections are displayed:

Optimization summary: If the minimum number of experiments is strictly inferior to the maximum number of experiments, then a table with information for each number of experiments is displayed. This table displays for each optimization run the number of experiments, the criterion $log(determinant)$, the criterion $norm.log(determinant)$ and the criterion $Log(|I|^{1/p})$. The best result is displayed in bold in the first line. The criterion $norm.log(determinant)$ is shown in a chart.

Statistics for each iteration: This table shows for the selected experimental design the evolution of the criterion during the iterations of the optimization. If the corresponding option is activated in the Charts tab, a chart showing the evolution of the criterion is displayed.

If the generation of **experiment sheets** was activated in the dialog box and if there are less than 200 experiments to be carried out, an experiment sheet is generated for each line of the experimental design on separate Excel sheets.

These sheets start with the report header of the experimental design and the model name to simplify the identification of the experimental design that this sheet belongs to. Then the running number of the experiment and the total number of experiments are displayed. The values of the additional columns of the experimental design, i. e. sort order, run order, and repetition are given for the experiment.

Last, the information on the experimental conditions of the factors is displayed with fields so that the user can enter the results obtained for the various responses. Short names, long names, units, physical units and values are displayed for each factor.

These sheets can be printed out or can be used in electronic format to assist during the realization of the experiments.

Example

A tutorial on the generation and analysis of a screening design is available on the XLSTAT Help Center:

http://www.xlstat.com/demo-doe1.htm

References

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 edition, John Wiley & Sons.

Myers, R. H., Khuri, I. K. and Carter W. H. Jr. (1989). Response Surface Methodology: 1966 – 1988, *Technometrics*, **31**, 137-157.

Analysis of a screening design

Use this tool to analyze a screening design of 2 to 35 factors and a user defined number of results. A linear model with or without interactions will be used for the analysis

In this section:

Description

Analysis of a screening design uses the same conceptual framework as linear regression and variance (ANOVA). The main difference comes from the nature of the underlying model. In ANOVA, explanatory variables are often called factors.

If p is the number of factors, the ANOVA model is written as follows

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_{k(i,j),j} + \epsilon_i \quad (1)$$

where Y_i is the value observed for the dependent variable for observation i, $k(i,j)$ is the index of the category of factor j for observation i, and ϵ_i is the error of the model.

The hypotheses used in ANOVA are identical to those used in linear regression: the errors ϵ_i follow the same normal distribution $\mathcal{N}(0,s)$ and are independent.

The way the model with this hypothesis added is written means that, within the framework of the linear regression model, the Y_is are the expression of random variables with mean μ_i and variance s^2, where

$$\mu_i = \beta_0 + \sum_{j=1}^{p} \beta_{k(i,j),j}$$

To use the various tests proposed in the results of linear regression, it is recommended to check retrospectively that the underlying hypotheses have been correctly verified. The normality of the residuals can be checked by analyzing certain charts or by using a normality test. The independence of the residuals can be checked by analyzing certain charts or by using the Durbin Watson test.

For more information on ANOVA and linear regression please consider the corresponding sections in the online help.

Responses optimization and desirability

In the case of many response values $y_1, ..., y_m$ it is possible to optimize each response value individually and to create a combined desirability function and analyze its values. Proposed by Derringer and Suich (1980), this approach is to first convert each response y_i into an individual desirability function d_i that varies over the range $0 \leq d_i \leq 1$.

When y_i has reached its target, then $d_i = 1$. If y_i is outside an acceptable region around the target, then $d_i = 0$. Between these two extreme cases, intermediate values of d_i exist as shown below.

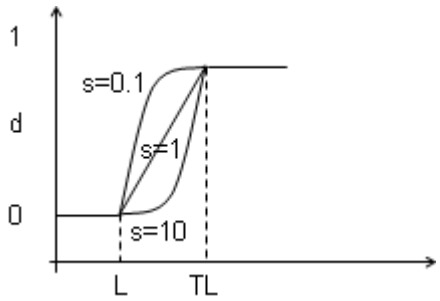
The 3 different optimization cases for d_i are present with the following definitions:

- L = lower value. Every value smaller than L has $d_i = 0$
- U = upper value. Every value bigger than U has $d_i = 0$.
- $T(L)$ = left target value.
- $T(R)$ = right target value. Every value between $T(L)$ and $T(R)$ has $d_i = 1$.
- s, t = weighting parameters that define the shape of the optimization function between L and $T(L)$ and $T(R)$ and U.

The following equation has to be respected when defining $L, U, T(L)$ and $T(R)$:

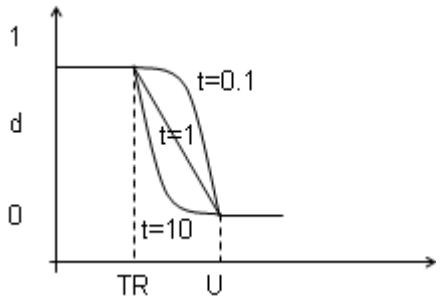
$$L \leq T(L) \leq T(R) \leq U$$

Maximize the value of y_i:



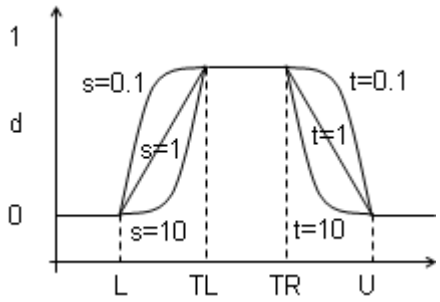
$$d_i = \begin{cases} 0 & y_i < L \\ (\frac{y_i - L}{T(L) - L})^s & L \leq y_i \leq T(L) \\ 1 & y_i > T(L) \end{cases}$$

Minimize the value of y_i:



$$d_i = \begin{cases} 1 & y_i < T(R) \\ (\frac{U-y_i}{U-T(R)})^t & T(R) \leq y_i \leq U \\ 0 & y_i > U \end{cases}$$

Two sided desirability function as shown below to **target** a certain interval of y_i:



$$d_i = \begin{cases} 0 & y_i < L \\ (\frac{y_i-L}{T(L)-L})^s & L \leq y_i \leq T(L) \\ 1 & T(L) < y_i < T(R) \\ (\frac{U-y_i}{U-T(R)})^t & T(R) \leq y_i \leq U \\ 0 & y_i > U \end{cases}$$

The design variables are chosen to maximize the overall desirability D

$$D = (d_1^{w_1} \cdot d_2^{w_2} \cdot \ldots \cdot d_m^{w_m})^{\frac{1}{w_1 \cdot w_2 \cdot \ldots \cdot w_m}}$$

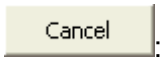
Where $1 \leq w_i \leq 10$ are weightings of the individual desirability functions. The bitter w_i, the more important is d_i taken into account during the optimization.

In the display, XLSTAT gives the 5 best solutions found during the optimization.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select.

General tab:

Y / results: Select the columns of the experimental design that contain the results. These columns should now hold the results of the experiments carried out. If several result variables have been selected, XLSTAT carries out calculations for each of the variables separately, and then an analysis of the desirability is carried out.

Experimental design: Select your experimental design. If you have changed your design, check that the qualitative and quantitative factors follow each other. All the columns of the design must be selected.

Number of quantitative factors: Enter the number of quantitative factors in your experimental design.

Number of qualitative factors: Enter the number of qualitative factors in your experimental design.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: This option is always activated. The first row of the selected data (data and observation labels) must contain a label.

Responses tab:

Optimization of responses: Activate this option if you wish to optimize responses. In this case select the responses optimization table generated when the design was created. The header of

the table must be included in the selection.

- **Objective**: Choose the objective of the optimization. You have the choice between minimum, optimum and maximum.

If the selected objective is the optimum or the maximum, the following fields are activated:

- **Lower**: Enter for each answer the value of the lower bound below which the desirability is 0.
- **Target (left)**: Enter the value of the lower bound above which desirability is 1 for each response.

If the selected objective is the optimum or the minimum, the following fields are activated:

- **Target (right)**: Enter for each response the value of the upper bound below which the desirability is equal to 1.
- **Lower**: Enter for each answer the value of the upper limit above which the desirability is 0.
- **s**: Activate this option if the increasing desirability function must be non-linear. Then enter the value of the shape parameter which must be between 0.01 and 100.
- **t**: Activate this option if the decreasing desirability function must be non-linear. Then enter the value of the shape parameter which must be between 0.01 and 100.
- **Weight**: Activate this option if the answers must have an exponential value different from 1 when calculating desirability. Then enter the value of the shape parameter which must be between 0.01 and 100.

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Outputs tab:

Analysis of variance: Activate this option to display the analysis of variance table.

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Contribution: Activate this option to display the contribution of the factors to the model.

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

- **Adjusted predictions**: Activate this option to calculate and display adjusted predictions in the table of predictions and residuals.
- **Cook's D**: Activate this option to calculate and display Cook's distances in the table of predictions and residuals.

- **Studentized residues**: Activate this option to calculate and display studentized residuals in the table of predictions and residuals.

Means: Activate this option to compute and display the means for the categories of the main and interaction factors.

- **LS Means**: Activate this option to compute least squares means instead of observed means.
- **Standard errors**: Activate this option to display the standard errors with the means
- **Confidence intervals**: Activate this option to additionally display the confidence intervals around the means.

Charts tab:

Regression charts: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.
- **Predictions and residuals**: Activate this option to display the following charts.

(1) Line of regression: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

(2) Explanatory variable versus standardized residuals: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

(3) Dependent variable versus standardized residuals.

(4) Predictions for the dependent variable versus the dependent variable.

(5) Bar chart of standardized residuals.

Confidence intervals: Activate this option to have confidence intervals displayed on charts (1) and (4).

Pareto plots: Activate this option, to display the chart that represents the contribution of the factors to the response in a Pareto plot.

Means charts: Activate this option to display the charts used to display the means of the various categories of the various factors.

Results

Variables information: This table shows the information about the factors. For each factor the short name, long name, unit and physical unit are displayed.

Responses optimization: This table gives the 5 best solutions obtained during the responses optimization.

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.
- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).
- **R²**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2} \text{ with } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i,$$

The R^2 is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer R^2 is to 1, the better is the model. The problem with the R^2 is that it does not take into account the number of variables used to fit the model.

- **Adjusted R²**: The adjusted determination coefficient for the model. The adjusted R^2 can be negative if the R^2 is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

$$\hat{R}^2 = 1 - (1 - R^2) \frac{W - 1}{W - p - 1}$$

The adjusted R^2 is a correction to the R^2 which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.
- **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows:

$$MAPE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- **DW**: The Durbin-Watson statistic is defined by:

$$DW = \frac{\sum_{i=2}^{n}[(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2}$$

This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- **Cp**: Mallows Cp coefficient is defined by:

$$Cp = \frac{SSE}{\hat{\sigma}} + 2p^* - W$$

where SSE is the sum of the squares of the errors for the model with p explanatory variables and $\hat{\sigma}$ is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to p^*, the less the model is biased.

- **AIC**: Akaike's Information Criterion is defined by:

$$AIC = W \ln(\frac{SSE}{W}) + 2p^*$$

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz's Bayesian Criterion is defined by:

$$SBC = W \ln(\frac{SSE}{W}) + \ln(W)p^*$$

This criterion, proposed by Schwarz (1978) is similar to the AIC, and the aim is to minimize it.

- **PC**: Amemiya's Prediction Criterion is defined by:

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}$$

This criterion, proposed by Amemiya (1980) is used, like the adjusted R^2 to take account of the parsimony of the model.

- **Press RMSE**: Press' statistic is only displayed if the corresponding option has been activated in the dialog box. It is defined by:

$$Press = \sum_{i=1}^{n} w_i(y_i - \hat{y}_{i(-i)})^2$$

where $\hat{y}_{i(-i)}$is the prediction for observation i when the latter is not used for estimating parameters. We then get:

$$\text{Press RMCE } = \sqrt{\frac{\text{Press}}{W - p^*}}$$

Press's RMSE can then be compared to the RMSE. A large difference between the two shows that the model is sensitive to the presence or absence of certain observations in the model.

- **Q²**: The Q^2 statistic is displayed. It is defined as

$$Q^2 = 1 - \frac{PressRMSE}{SSE}$$

The closer Q^2 is to 1, the better and more robust is the model.

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

If the Type I/II/III SS (SS: Sum of Squares) is activated, the corresponding tables are displayed.

The table of **Type I SS** values is used to visualize the influence that progressively adding explanatory variables has on the fitting of the model, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. The sums of squares in the Type I table always add up to the model SS. Note: the order in which the variables are selected in the model influences the values obtained.

The table of **Type II SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained.

The table of **Type III SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, expect those were the effect is present (interactions), as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. Type II and Type III are identical if there are no interactions or if the design is balanced.

The **parameters of the model** table displays the estimate of the parameters, the corresponding standard error, the Student's t, the corresponding probability, as well as the confidence interval

The **equation of the model** is then displayed to make it easier to read or re-use the model.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the residuals, the confidence intervals together with the fitted prediction and Cook's D if the corresponding options have been activated in the dialog box. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the random values being larger.

The **charts** which follow show the results mentioned above. If there is only one explanatory variable in the model, the first chart displayed shows the observed values, the regression line and both types of confidence interval around the predictions. The second chart shows the standardized residuals as a function of the explanatory variable. In principle, the residuals should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a problem with the model.

The **three charts** displayed next respectively show the evolution of the standardized residuals as a function of the dependent variable, the distance between the predictions and the observations (for an ideal model, the points would all be on the bisector), and the standardized residuals on a bar chart. The last chart quickly shows if an abnormal number of values are outside the interval $]-2,2[$ given that the latter, assuming that the sample is normally distributed, should contain about 95% of the data.

Example

A tutorial on the generation and the analysis of a screening design is available on the XLSTAT Help Center:

http://www.xlstat.com/demo-doe3.htm

References

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 edition, John Wiley & Sons.

Myers, R. H., Khuri, I. K. and Carter W. H. Jr. (1989). Response Surface Methodology: 1966 – 1988, *Technometrics*, **31**, 137-157.

Surface response designs

Use this module to generate a design to analyze the surface response for 2 to 10 factors and one or more responses.

In this section:

Description

The family of surface response design is used for modeling and analysis of problems in which a response of interest is influenced by several variables and the objective is to optimize this response.

Remark: In contrast to this, screening designs aim to study the input factors, not the response value.

For example, suppose that an engineer wants to find the optimal levels of the pressure (x_1) and the temperature (x_2) of an industrial process to produce concrete, which should have a maximum hardness y.

$$y = f(x_1, x_2) + \epsilon_i \quad (1)$$

Model

This tool supposes a second-order model. If k is the number of factors, the quadratic model is written as follows:

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j} \sum \beta_{ij} x_i x_j + \epsilon \quad (2)$$

Design

The tool offers the following design approaches for surface modeling:

Full factorial design with 2 levels: All combinations of 2 values for each factor (minimum and maximum) are generated in the design. The number of experiments n for k factors is given as:

$$n = 2^k$$

Full factorial design with 3 levels: All combinations of 3 values for each factor (minimum, mean and maximum) are generated in the design. The number of experiments n for k factors is given as:

$$n = 3^k$$

Central composite design: Proposed by Box G.E.P. and Wilson K.B. (1951), the points of experiments are generated on a sphere around the center point. The number of different factor levels is minimized. The center point is repeated in order to maximize the prediction precision around the supposed optimum. The number of repetitions n_0 of the center point is calculated by the following formulas for k factors based on the uniform precision:

$$\gamma = \frac{(k+3) + \sqrt{9k^2 + 14k - 7}}{4(k+2)}$$

and

$$n_0 = floor(\gamma(\sqrt{2}+2)^2 - 2^k - 2k)$$

, where floor designates the biggest integer value smaller than the argument. The number of experiments n for k factors is given as:

$$n = 2^k + 2k + 1$$

Box-Behnken: This design was proposed by Box G.E.P. and Behnken D.W (1960) and is based on the same principles as the central composite design, but with a smaller number of experiments. The number of experiments n for k factors is given as:

$$n = 2k^2 - 2k + 1$$

Doehlert: This design was proposed by Doehlert D.H. (1970) and is based on the same principles as the central composite and Box-Behnken design, but with a smaller number of experiments. This design has a larger amount of different factor levels for several factors of the design and might therefore be difficult to use. The number of experiments n for k factors is given as:

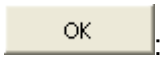
$$n = k^2 + k + 1$$

The following table displays the number of different experiments for each of the 4 design choices and a given number of factors k to be analyzed. In this calculation, the center point is only present one time.

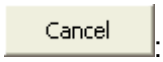
k	full fact.	Cent. comp.	Box-Behnken	Doehlert
2	9	9	5	7
3	27	15	13	13
4	81	25	25	21
5	243	43	41	31
6	729	77	61	43
7	2187	143	85	57

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.


: Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select.

General tab:

Quantitative factors (min/max): select the minimum and the maximum of the quantitative factors. The selection must have two rows, the first corresponds to the minimums, the second to the maximums for each of the factors.

Qualitative factors: select the qualitative factors and their modalities.

Number of responses: enter the number of responses to analyze.

Repetitions: Activate this option to choose the number of repetitions of the design.

Randomize: Activate this option to change the order of the lines of the design into a random order.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections includes a header.

Options tab:

Design of experiments: Choose the experiment plan you want to use. Depending on the number of factors, different plans are offered.

Number of central points: In the case of a centered composite design, you have the possibility to change the number of repetitions of the central point. Enable this option to enter a value for this number.

Outputs tab:

Coded design: Activate this option to display the table of the encoded design of experiments.

Display experiments sheets: Activate this option if you want to display individual Excel sheets for each experience. This can be useful in order to print them out and carry out the experiments.

Results

Results

Variables information: This table shows the information about the factors. For each factor the short name, long name, unit and physical unit are displayed.

Experimental design: This table displays the complete experimental design. Additional columns include information on the factors and on the responses, a label for each experiment, the sort order, the run order and the repetition.

Responses optimization: The responses optimization table of is displayed after the experimental design. You must then select the following parameters:

- **Objective**: Choose the objective of the optimization. You have the choice between minimum, optimum and maximum.

If the selected objective is the optimum or the maximum, the following fields are activated:

- **Lower**: Enter for each answer the value of the lower bound below which the desirability is 0.

- **Target (left)**: Enter the value of the lower bound above which desirability is 1 for each response.

If the selected objective is the optimum or the minimum, the following fields are activated:

- **Target (right)**: Enter for each response the value of the upper bound below which the desirability is equal to 1.
- **Lower**: Enter for each answer the value of the upper limit above which the desirability is 0.
- **s**: Activate this option if the increasing desirability function must be non-linear. Then enter the value of the shape parameter which must be between 0.01 and 100.
- **t**: Activate this option if the decreasing desirability function must be non-linear. Then enter the value of the shape parameter which must be between 0.01 and 100.
- **Weight**: Activate this option if the answers must have an exponential value different from 1 when calculating desirability. Then enter the value of the shape parameter which must be between 0.01 and 100.

For more details on responses optimization, you can refer to the analysis of a factor effect design.

Endoded design ; This table shows the encoded experimental design. This table is only displayed in the case of a d-optimal experimental design.

Example

A tutorial on the generation and analysis of a surface response design is available on the XLSTAT Help Center:

http://www.xlstat.com/demo-doe2.htm

References

Box G. E. P. and Behnken D. W. (19 60). 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. (19 51). On the experimental attainment of optimum conditions, *Journal of Royal Statistical Society*, **13**, Serie B, 1-45.

Doehlert D. H. (19 70). 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 edition, John Wiley & Sons.

Myers, R. H., Khuri, I. K. and Carter W. H. Jr. (1989). Response Surface Methodology: 1966 – 1988, *Technometrics*, **31**, 137-157.

Analysis of a Surface response design

Use this tool to analyze a surface response design for 2 to 10 factors and a user defined number of results. A second order model is used for the analysis.

In this section:

Description

The analysis of a surface response design uses the same statistical and conceptual framework as linear regression. The main difference comes from the model that is used. A quadratic form is used as a model

If k is the number of factors, the quadratic model is written as follows:

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j} \sum \beta_{ij} x_i x_j + \epsilon \qquad (1)$$

For more information on ANOVA and linear regression please consider the corresponding sections in the online help.

Responses optimization and desirability

In the case of many response values $y_1, ..., y_m$ it is possible to optimize each response value individually and to create a combined desirability function and analyze its values. Proposed by Derringer and Suich (1980), this approach is to first convert each response y_i into an individual desirability function d_i that varies over the range $0 \leq d_i \leq 1$.

When y_i has reached its target, then $d_i = 1$. If y_i is outside an acceptable region around the target, then $d_i = 0$. Between these two extreme cases, intermediate values of d_i exist as shown below.

The 3 different optimization cases for d_i are present with the following definitions:

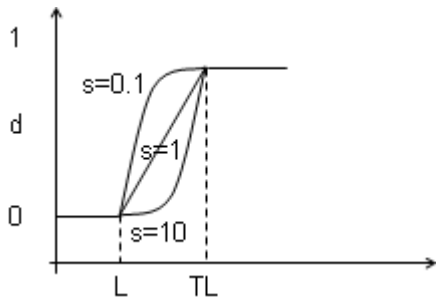
- L = lower value. Every value smaller than L has $d_i = 0$
- U = upper value. Every value bigger than U has $d_i = 0$.

- $T(L)$ = left target value.
- $T(R)$ = right target value. Every value between $T(L)$ and $T(R)$ has $d_i = 1$.
- s, t = weighting parameters that define the shape of the optimization function between L and $T(L)$ and $T(R)$ and U.

The following equation has to be respected when defining $L, U, T(L)$ and $T(R)$:

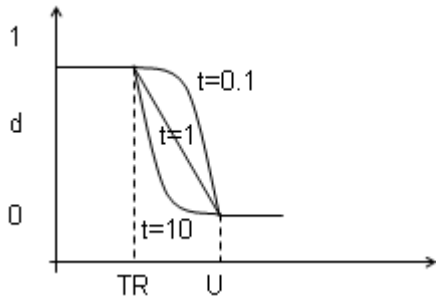
$$L \leq T(L) \leq T(R) \leq U$$

Maximize the value of y_i:



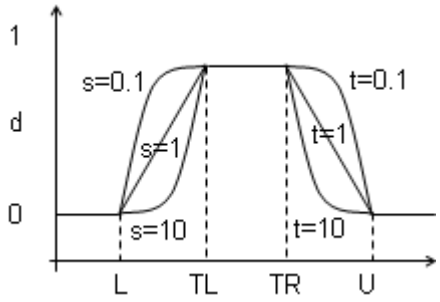
$$d_i = \begin{cases} 0 & y_i < L \\ (\frac{y_i - L}{T(L) - L})^s & L \leq y_i \leq T(L) \\ 1 & y_i > T(L) \end{cases}$$

Minimize the value of y_i:



$$d_i = \begin{cases} 1 & y_i < T(R) \\ (\frac{U - y_i}{U - T(R)})^t & T(R) \leq y_i \leq U \\ 0 & y_i > U \end{cases}$$

Two sided desirability function as shown below to **target** a certain interval of y_i:



$$d_i = \begin{cases} 0 & y_i < L \\ (\frac{y_i - L}{T(L) - L})^s & L \leq y_i \leq T(L) \\ 1 & T(L) < y_i < T(R) \\ (\frac{U - y_i}{U - T(R)})^t & T(R) \leq y_i \leq U \\ 0 & y_i > U \end{cases}$$

The design variables are chosen to maximize the overall desirability D

$$D = (d_1^{w_1} \cdot d_2^{w_2} \cdot \ldots \cdot d_m^{w_m})^{\frac{1}{w_1 \cdot w_2 \cdot \ldots \cdot w_m}}$$

Where $1 \leq w_i \leq 10$ are weightings of the individual desirability functions. The bitter w_i, the more important is d_i taken into account during the optimization.

In the display, XLSTAT gives the 5 best solutions found during the optimization.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

: Click this button to reload the default options.

: Click this button to delete the data selections.

: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects that you select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select.

General tab:

Y / results: Select the columns of the experimental design that contain the results. These columns should now hold the results of the experiments carried out. If several result variables have been selected, XLSTAT carries out calculations for each of the variables separately, and then an analysis of the desirability is carried out.

Experimental design: Select your experimental design. If you have changed your design, check that the qualitative and quantitative factors follow each other. All the columns of the design must be selected.

Number of quantitative factors: Enter the number of quantitative factors in your experimental design.

Number of qualitative factors: Enter the number of qualitative factors in your experimental design.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: This option is always activated. The first row of the selected data (data and observation labels) must contain a label.

Responses tab:

Optimization of responses: Activate this option if you wish to optimize responses. In this case select the responses optimization table generated when the design was created. The header of the table must be included in the selection.

- **Objective**: Choose the objective of the optimization. You have the choice between minimum, optimum and maximum.

If the selected objective is the optimum or the maximum, the following fields are activated:

- **Lower**: Enter for each answer the value of the lower bound below which the desirability is 0.
- **Target (left)**: Enter the value of the lower bound above which desirability is 1 for each response.

If the selected objective is the optimum or the minimum, the following fields are activated:

- **Target (right)**: Enter for each response the value of the upper bound below which the desirability is equal to 1.
- **Lower**: Enter for each answer the value of the upper limit above which the desirability is 0.
- **s**: Activate this option if the increasing desirability function must be non-linear. Then enter the value of the shape parameter which must be between 0.01 and 100.
- **t**: Activate this option if the decreasing desirability function must be non-linear. Then enter the value of the shape parameter which must be between 0.01 and 100.
- **Weight**: Activate this option if the answers must have an exponential value different from 1 when calculating desirability. Then enter the value of the shape parameter which must be between 0.01 and 100.

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Outputs tab:

Analysis of variance: Activate this option to display the analysis of variance table.

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Contribution: Activate this option to display the contribution of the factors to the model.

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

- **Adjusted predictions**: Activate this option to calculate and display adjusted predictions in the table of predictions and residuals.
- **Cook's D**: Activate this option to calculate and display Cook's distances in the table of predictions and residuals.
- **Studentized residues**: Activate this option to calculate and display studentized residuals in the table of predictions and residuals.

Means: Activate this option to compute and display the means for the categories of the main and interaction factors.

- **LS Means**: Activate this option to compute least squares means instead of observed means.
- **Standard errors**: Activate this option to display the standard errors with the means

- **Confidence intervals**: Activate this option to additionally display the confidence intervals around the means.

Charts tab:

Regression charts: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.
- **Predictions and residuals**: Activate this option to display the following charts.

(1) Line of regression: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

(2) Explanatory variable versus standardized residuals: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

(3) Dependent variable versus standardized residuals.

(4) Predictions for the dependent variable versus the dependent variable.

(5) Bar chart of standardized residuals.

Confidence intervals: Activate this option to have confidence intervals displayed on charts (1) and (4).

Contour plot: Activate this option to display charts the represent the desirability function in contour plots in the case of a model with 2 factors.

Trace plot: Activate this option to display charts the represent the trace of the desirability function for each of the factors, with the other factors set to the mean value.

Results

Variables information: This table shows the information about the factors. For each factor the short name, long name, unit and physical unit are displayed.

Responses optimization: This table gives the 5 best solutions obtained during the responses optimization.

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.
- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).

- **R²**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2} \text{ with } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i,$$

The R^2 is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer R^2 is to 1, the better is the model. The problem with the R^2 is that it does not take into account the number of variables used to fit the model.

- **Adjusted R²**: The adjusted determination coefficient for the model. The adjusted R^2 can be negative if the R^2 is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

$$\hat{R}^2 = 1 - (1 - R^2) \frac{W - 1}{W - p - 1}$$

The adjusted R^2 is a correction to the R^2 which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

- **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows:

$$MAPE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- **DW**: The Durbin-Watson statistic is defined by:

$$DW = \frac{\sum_{i=2}^{n} [(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}$$

This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- **Cp**: Mallows Cp coefficient is defined by:

$$Cp = \frac{SSE}{\hat{\sigma}} + 2p^* - W$$

where SSE is the sum of the squares of the errors for the model with p explanatory variables and $\hat{\sigma}$ is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to p^*, the less the model is biased.

- **AIC**: Akaike's Information Criterion is defined by:

$$AIC = W \ln(\frac{SSE}{W}) + 2p^*$$

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz's Bayesian Criterion is defined by:

$$SBC = W \ln(\frac{SSE}{W}) + \ln(W)p^*$$

This criterion, proposed by Schwarz (1978) is similar to the AIC, and the aim is to minimize it.

- **PC**: Amemiya's Prediction Criterion is defined by:

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}$$

This criterion, proposed by Amemiya (1980) is used, like the adjusted R^2 to take account of the parsimony of the model.

- **Press RMSE**: Press' statistic is only displayed if the corresponding option has been activated in the dialog box. It is defined by:

$$Press = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i(-i)})^2$$

where $\hat{y}_{i(-i)}$is the prediction for observation i when the latter is not used for estimating parameters. We then get:

$$\text{Press RMCE } = \sqrt{\frac{\text{Press}}{W - p^*}}$$

Press's RMSE can then be compared to the RMSE. A large difference between the two shows that the model is sensitive to the presence or absence of certain observations in the model.

- **Q²**: The Q^2 statistic is displayed. It is defined as

$$Q^2 = 1 - \frac{PressRMSE}{SSE}$$

The closer Q^2 is to 1, the better and more robust is the model.

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

If the Type I/II/III SS (SS: Sum of Squares) is activated, the corresponding tables are displayed.

The table of **Type I SS** values is used to visualize the influence that progressively adding explanatory variables has on the fitting of the model, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. The sums of squares in the Type I table always add up to the model SS. Note: the order in which the variables are selected in the model influences the values obtained.

The table of **Type II SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained.

The table of **Type III SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, expect those were the effect is present (interactions), as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. Type II and Type III are identical if there are no interactions or if the design is balanced.

The **parameters of the model** table displays the estimate of the parameters, the corresponding standard error, the Student’s t, the corresponding probability, as well as the confidence interval

The **equation of the model** is then displayed to make it easier to read or re-use the model.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the residuals, the confidence intervals together with the fitted prediction and Cook's D if the corresponding options have been activated in the dialog box. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the random values being larger.

The **charts** which follow show the results mentioned above. If there is only one explanatory variable in the model, the first chart displayed shows the observed values, the regression line and both types of confidence interval around the predictions. The second chart shows the standardized residuals as a function of the explanatory variable. In principle, the residuals should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a problem with the model.

The **three charts** displayed next respectively show the evolution of the standardized residuals as a function of the dependent variable, the distance between the predictions and the observations (for an ideal model, the points would all be on the bisector), and the standardized residuals on a bar chart. The last chart quickly shows if an abnormal number of values are outside the interval $]-2,2[$ given that the latter, assuming that the sample is normally distributed, should contain about 95% of the data.

Then the **contour plot** is displayed, if the design has two factors and the corresponding option is activated. The contour plot is shown as a two dimensional projection and as a 3D chart. Using these charts it is possible to analyze the dependence of the two factors simultaneously.

Then the **trace plots** are displayed, if the corresponding option is activated. The trace plots show for each factor the response variable as a function of the factor. All other factors are set to their mean value. These charts are shown in two options: with the standardized factors and with the factors in original values. Using these plots the dependence of a response on a given factor can be analyzed.

Example

A tutorial on the generation and the analysis of a surface response design is available on the XLSTAT Help Center:

http://www.xlstat.com/demo-doe2.htm

References

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 edition, John Wiley & Sons.

Myers, R. H., Khuri, I. K. and Carter W. H. Jr. (1989). Response Surface Methodology: 1966 – 1988, *Technometrics*, **31**, 137-157.

Mixture designs

Use this module to generate a mixture design for 2 to 6 factors.

In this section:

Description

Mixture designs are used to model the results of experiments where these relate to the optimization of formulation. The resulting model is called "mixture distribution."

Mixture designs differ from factorial designs because of the following characteristics:

- The factors studied are proportions whose sum is equal to 1.
- Construction of experimental designs is subject to constraints because the factors may not evolve independently of each other (the sum of the proportions being 1).

Experimental space of a mixture

When the concentrations of the n components are not submitted to any constraints, the experimental design is a simplex, that is to say, a regular polyhedron with n vertices in a space of dimension $n-1$. For example, for a mixture of three components, the experimental field is an equilateral triangle; for 4 constituents it is a regular tetrahedron.

Creating mixture designs therefore consists of positioning the experiments regularly in the simplex to optimize the accuracy of the model. The most conventional designs are Scheffé's designs, Scheffé-centroid designs, and augmented designs.

If constraints on the components of the model are introduced by defining a minimum amount or a maximum amount not to exceed, then the experimental domain can be a simplex, an inverted simplex (also called simplex B) or any convex polyhedron. In the latter case, the simplex designs are no longer usable.

Warning: if there is a large number of components and there are many constraints on the components, it is possible that the experimental domain would not exist.

The Scheffé simplex networks are the easiest designs to build. They allow you to build models of any degree m. These matrices are related to a canonical model having a high number of coefficients (Full Canonical Model).

||**Degree of the model** ||| ---|---|---|--- **Constituents** | **2** | **3** | **43** | 6 | 10 | 15
4 | 10 | 20 | 35
5 | 15 | 35 | 70
6 | 21 | 56 | 126
8 | 36 | 120 | 330
10 | 55 | 220 | 715

To improve the sequentiality of the experiments, Scheffé proposed adding points to the center of experimental space. These experimental designs are known Simplex Centroid Designs.

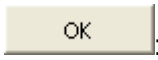
These mixture designs allow you to construct a reduced polynomial model, which comprises only product terms of the components. The number of experiments thus increases less rapidly than in the case of a Scheffé's simplex. Centered simplexes add additional mixtures in the center of the experimental space compared with conventional simplexes. This has the effect of improving the quality of predictions in the center of the field.

Output

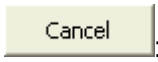
This tool will provide a new design for testing. Optional experiment sheets for each individual test might be generated on separated Excel sheets for printing. After carrying out the experiments, complete the corresponding cells in the created experimental design in the corresponding Excel sheet.

Dialog box

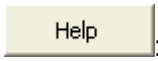
The dialog box is divided into several tabs that correspond to a variety of options, ranging from the selection of data to the display of results. Below are descriptions of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

 : Click these buttons to change the data selection mode. If the button shows a mouse icon, XLSTAT expects you to select data with the mouse. If the button shows a list icon, XLSTAT will display a list of the available variables that you can select.

General tab:

Quantitative factors (min/max): select the minimum and the maximum of the quantitative factors. The selection must have two rows, the first corresponds to the minimums, the second to the maximums for each of the factors.

Number of responses: enter the number of responses to analyze.

Repetitions: Activate this option to choose the number of repetitions of the design.

Randomize: Activate this option to change the order of the lines of the design to a random order.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections includes a header.

Options tab:

Design of experiments: Choose the experiment plan you want to use. Depending on the number of factors, different plans are offered.

Degree of the model: In the case of a simplex design, you can choose the number of degrees of the model.

Total quantity of mixture: Enter the total quantity of mixture. This is the quantity of mixture used in the experiment.

Outputs tab:

Display experiments sheets: Activate this option if you want to display individual Excel sheets for each experiment. This can be useful for printingg them out before carrying out the experiments.

Results

Variables information: This table shows the information about the factors.

Experimental design: This table displays the complete experimental design. Additional columns include information on the factors and on the responses, a label for each experiment, the sort order, the run order, and the repetition.

Responses optimization: The responses optimization table of is displayed after the experimental design. You must then select the following parameters:

- **Objective**: Choose the objective of the optimization. You have the choice between minimum, optimum and maximum.

If the selected objective is the optimum or the maximum, the following fields are activated:

- **Lower**: Enter for each answer the value of the lower bound below which the desirability is 0.
- **Target (left)**: Enter the value of the lower bound above which desirability is 1 for each response.

If the selected objective is the optimum or the minimum, the following fields are activated:

- **Target (right)**: Enter for each response the value of the upper bound below which the desirability is equal to 1.
- **Lower**: Enter for each answer the value of the upper limit above which the desirability is 0.
- **s**: Activate this option if the increasing desirability function must be non-linear. Then enter the value of the shape parameter, which must be between 0.01 and 100.
- **t**: Activate this option if the decreasing desirability function must be non-linear. Then enter the value of the shape parameter which must be between 0.01 and 100.
- **Weight**: Activate this option if the answers must have an exponential value different from 1 when calculating desirability. Then enter the value of the shape parameter, which must be between 0.01 and 100.

For more details on responses optimization, you can refer to the analysis of a screening design.

Example

A tutorial on the generation and analysis of a mixture design is available on the XLSTAT Help Center:

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

References

Droesbeke J.J., Fine J. and Saporta G. (1997). Plans d'Expériences - Application Industrielle. Editions Technip.

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.

Louvet F. and Delplanque L. (2005). Design Of Experiments: The French touch, Les plans d’expériences : une approche pragmatique et illustrée, Alpha Graphic, Olivet.

Analysis of a mixture design

Use this tool to analyze a mixture design for 2 to 6 factors.

In this section:

Description

The analysis of a mixture design is based on the same principle as linear regression. The major difference comes from the model that is used. Several models are available.

By default, XLSTAT associates a reduced model (Simplified Canonical Model) to centroïd simplexes. However, it is possible to change the model if the number of degrees of freedom is sufficient (by increasing the number of repetitions of the experiments).

To fulfil the constraint associated to a mixture design, a polynomial model with no intercept is used. We distinguish two types of models, simplified (special) models and full models (from level 3).

The model equations are:

- Linear model (level 1):

$$Y = \sum_i \beta_i x_i$$

- Quadratic model (level 2):

$$Y = \sum_i \beta_i x_i + \sum_i \sum_{i<j} \beta_{ij} x_i x_j$$

- Full cubic model (level 3):

$$Y = \sum_i \beta_i x_i + \sum_i \sum_{i<j} \beta_{ij} x_i x_j + \sum_j \sum_{i<j} \delta_{ij} x_i x_j (x_i - x_j) + \sum_k \sum_{j<k} \beta_{ijk} x_i x_j (x_k)$$

- Simplified cubic model (special):

$$Y = \sum_i \beta_i x_i + \sum_i \sum_{i<j} \beta_{ij} x_i x_j + \sum_k \sum_{j<k} \beta_{ijk} x_i x_j (x_k)$$

XLSTAT allows to apply models up to level 4 (simplified and full quartic model).

Estimation of these models is done with classical regression. For more details on ANOVA and linear regression, please refer to the chapters of this help associated to these methods.

Responses optimization and desirability

In the case of many response values $y_1, ..., y_m$ it is possible to optimize each response value individually and to create a combined desirability function and analyze its values. Proposed by Derringer and Suich (1980), this approach is to first convert each response y_i into an individual desirability function d_i that varies over the range $0 \leq d_i \leq 1$.

When y_i has reached its target, then $d_i = 1$. If y_i is outside an acceptable region around the target, then $d_i = 0$. Between these two extreme cases, intermediate values of d_i exist as shown below.

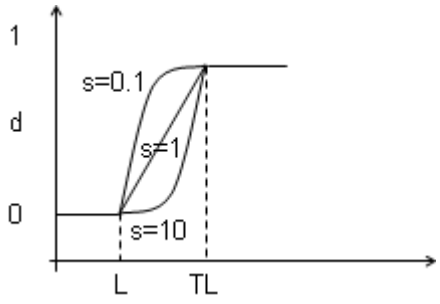
The 3 different optimization cases for d_i are present with the following definitions:

- L = lower value. Every value smaller than L has $d_i = 0$
- U = upper value. Every value bigger than U has $d_i = 0$.
- $T(L)$ = left target value.
- $T(R)$ = right target value. Every value between $T(L)$ and $T(R)$ has $d_i = 1$.
- s, t = weighting parameters that define the shape of the optimization function between L and $T(L)$ and $T(R)$ and U.

The following equation has to be respected when defining $L, U, T(L)$ and $T(R)$:

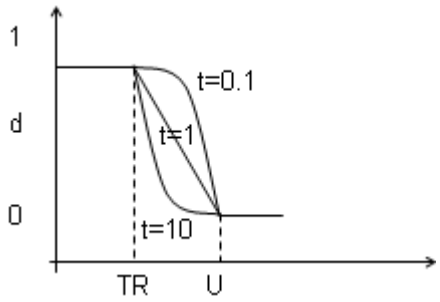
$$L \leq T(L) \leq T(R) \leq U$$

Maximize the value of y_i:



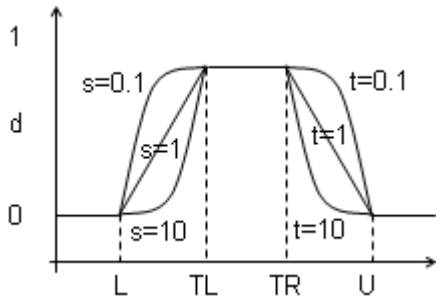
$$d_i = \begin{cases} 0 & y_i < L \\ (\frac{y_i - L}{T(L) - L})^s & L \leq y_i \leq T(L) \\ 1 & y_i > T(L) \end{cases}$$

Minimize the value of y_i:



$$d_i = \begin{cases} 1 & y_i < T(R) \\ (\frac{U-y_i}{U-T(R)})^t & T(R) \leq y_i \leq U \\ 0 & y_i > U \end{cases}$$

Two sided desirability function as shown below to **target** a certain interval of y_i:



$$d_i = \begin{cases} 0 & y_i < L \\ (\frac{y_i-L}{T(L)-L})^s & L \leq y_i \leq T(L) \\ 1 & T(L) < y_i < T(R) \\ (\frac{U-y_i}{U-T(R)})^t & T(R) \leq y_i \leq U \\ 0 & y_i > U \end{cases}$$

The design variables are chosen to maximize the overall desirability D

$$D = (d_1^{w_1} \cdot d_2^{w_2} \cdot \ldots \cdot d_m^{w_m})^{\frac{1}{w_1 \cdot w_2 \cdot \ldots \cdot w_m}}$$

Where $1 \leq w_i \leq 10$ are weightings of the individual desirability functions. The bitter w_i, the more important is d_i taken into account during the optimization.

In the display, XLSTAT gives the 5 best solutions found during the optimization.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / results: Select the columns of the experimental design that contain the results. These columns should now hold the results of the experiments carried out. If several result variables have been selected, XLSTAT carries out calculations for each of the variables separately, and then an analysis of the desirability is carried out.

Experimental design: Select your experimental design. If you have changed your design, check that the qualitative and quantitative factors follow each other. All the columns of the design must be selected.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: This option is always activated. The first row of the selected data (data and observation labels) must contain a label.

Responses tab:

Optimization of responses: Activate this option if you wish to optimize responses. In this case select the responses optimization table generated when the design was created. The header of the table must be included in the selection.

- **Objective**: Choose the objective of the optimization. You have the choice between minimum, optimum and maximum.

If the selected objective is the optimum or the maximum, the following fields are activated:

- **Lower**: Enter for each answer the value of the lower bound below which the desirability is 0.
- **Target (left)**: Enter the value of the lower bound above which desirability is 1 for each response.

If the selected objective is the optimum or the minimum, the following fields are activated:

- **Target (right)**: Enter for each response the value of the upper bound below which the desirability is equal to 1.
- **Lower**: Enter for each answer the value of the upper limit above which the desirability is 0.
- **s**: Activate this option if the increasing desirability function must be non-linear. Then enter the value of the shape parameter which must be between 0.01 and 100.
- **t**: Activate this option if the decreasing desirability function must be non-linear. Then enter the value of the shape parameter which must be between 0.01 and 100.
- **Weight**: Activate this option if the answers must have an exponential value different from 1 when calculating desirability. Then enter the value of the shape parameter which must be between 0.01 and 100.

Models: Select the type of model you want to use during your analysis. Depending on the model chosen, the factors corresponding to this model will be preselected when the second interface pops up. You can then select or deselect the factors you want to use.

Total quantity of mixture: Enter the total quantity of mixture. It is the quantity of mixture used in the experiment. It must match the one entered when creating the mixture design.

Outputs tab:

Correlations: Activate this option to display the correlation matrix for quantitative variables (dependent or explanatory).

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Analysis of variance: Activate this option to display the analysis of variance table.

Contribution: Activate this option to display the contribution of the factors to the model.

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

- **Adjusted predictions**: Activate this option to calculate and display adjusted predictions in the table of predictions and residuals.
- **Cook's D**: Activate this option to calculate and display Cook's distances in the table of predictions and residuals.
- **Studentized residues**: Activate this option to calculate and display studentized residuals in the table of predictions and residuals.

Charts tab:

Regression charts: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.

- **Predictions and residuals**: Activate this option to display the following charts.

(1) Line of regression: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

(2) Explanatory variable versus standardized residuals: This chart is only displayed if there is only one explanatory variable and this variable is quantitative.

(3) Dependent variable versus standardized residuals.

(4) Predictions for the dependent variable versus the dependent variable.

(5) Bar chart of standardized residuals.

Confidence intervals: Activate this option to have confidence intervals displayed on charts (1) and (4).

Ternary diagram: Activate this option to display a ternary diagram. This chart is displayed using 3 factors.

Results

Variables information: This table shows the information about the factors. For each factor the short name, long name, unit and physical unit are displayed.

Responses optimization: This table gives the 5 best solutions obtained during the responses optimization.

Correlation matrix: This table is displayed to give you a view of the correlations between the various variables selected.

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.

- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.

- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).

- **R²**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2} \text{ with } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i,$$

The R^2 is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer R^2 is to 1, the better is the model. The problem with the R^2 is that it does not take into account the number of variables used to fit the model.

- **Adjusted R²**: The adjusted determination coefficient for the model. The adjusted R^2 can be negative if the R^2 is near to zero. This coefficient is only calculated if the constant of the model has not been fixed by the user. Its value is defined by:

$$\hat{R}^2 = 1 - (1 - R^2) \frac{W - 1}{W - p - 1}$$

The adjusted R^2 is a correction to the R^2 which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.

- **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows:

$$MAPE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- **DW**: The Durbin-Watson statistic is defined by:

$$DW = \frac{\sum_{i=2}^{n} [(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}$$

This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- **Cp**: Mallows Cp coefficient is defined by:

$$Cp = \frac{SSE}{\hat{\sigma}} + 2p^* - W$$

where SSE is the sum of the squares of the errors for the model with p explanatory variables and $\hat{\sigma}$ is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to p^*, the less the model is biased.

- **AIC**: Akaike's Information Criterion is defined by:

$$AIC = W \ln(\frac{SSE}{W}) + 2p^*$$

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz's Bayesian Criterion is defined by:

$$SBC = W \ln(\frac{SSE}{W}) + \ln(W)p^*$$

This criterion, proposed by Schwarz (1978) is similar to the AIC, and the aim is to minimize it.

- **PC**: Amemiya's Prediction Criterion is defined by:

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}$$

This criterion, proposed by Amemiya (1980) is used, like the adjusted R^2 to take account of the parsimony of the model.

- **Press RMSE**: Press' statistic is only displayed if the corresponding option has been activated in the dialog box. It is defined by:

$$Press = \sum_{i=1}^{n} w_i(y_i - \hat{y}_{i(-i)})^2$$

where $\hat{y}_{i(-i)}$is the prediction for observation i when the latter is not used for estimating parameters. We then get:

$$\text{Press RMCE } = \sqrt{\frac{\text{Press}}{W - p^*}}$$

Press's RMSE can then be compared to the RMSE. A large difference between the two shows that the model is sensitive to the presence or absence of certain observations in the model.

- **Q²**: The Q^2 statistic is displayed. It is defined as

$$Q^2 = 1 - \frac{PressRMSE}{SSE}$$

The closer Q^2 is to 1, the better and more robust is the model.

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. Where the constant of the model is not set to a given value, the explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable. Where the constant of the model is set, the comparison is made with respect to the model for which the dependent variable is equal to the constant which has been set.

If the **Type I/II/III SS** (SS: Sum of Squares) is activated, the corresponding tables are displayed.

The table of **Type I SS** values is used to visualize the influence that progressively adding explanatory variables has on the fitting of the model, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. The sums of squares in the Type I table always add up to the model SS. Note: the order in which the variables are selected in the model influences the values obtained.

The table of **Type II SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained.

The table of **Type III SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, expect those were the effect is present (interactions), as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. Type II and Type III are identical if there are no interactions or if the design is balanced.

The **parameters of the model** table displays the estimate of the parameters, the corresponding standard error, the Student’s t, the corresponding probability, as well as the confidence interval.

The **equation of the model** is then displayed to make it easier to read or re-use the model.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the residuals, the confidence intervals together with the fitted prediction and Cook's D if the corresponding options have been activated in the dialog box. Two types of confidence interval are displayed: a confidence interval around the mean (corresponding to the case where the prediction would be made for an infinite number of observations with a set of given values for the explanatory variables) and an interval around the

isolated prediction (corresponding to the case of an isolated prediction for the values given for the explanatory variables). The second interval is always greater than the first, the random values being larger.

The **charts** which follow show the results mentioned above. If there is only one explanatory variable in the model, the first chart displayed shows the observed values, the regression line and both types of confidence interval around the predictions. The second chart shows the standardized residuals as a function of the explanatory variable. In principle, the residuals should be distributed randomly around the X-axis. If there is a trend or a shape, this shows a problem with the model.

The **three charts** displayed next respectively show the evolution of the standardized residuals as a function of the dependent variable, the distance between the predictions and the observations (for an ideal model, the points would all be on the bisector), and the standardized residuals on a bar chart. The last chart quickly shows if an abnormal number of values are outside the interval $]-2,2[$ given that the latter, assuming that the sample is normally distributed, should contain about 95% of the data.

For each combination of factors, we draw a **ternary diagram**. This graph shows a response surface on one of the faces of the polyhedron to which the experimental space corresponds. These charts facilitate the interpretation of the model and allow to identify the optimal configurations.

Example

A tutorial on the generation and the analysis of a mixture design is available on the XLSTAT Help Center:

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

References

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.

Taguchi designs

Use this tool to generate Taguchi designs to find products that are robust and insensitive to the natural variability of the environment and processes.

In this section:

Description

The Taguchi method is a method introduced by Genichi Taguchi (Genichi and Wu, 1980) which is a method of design of experiment providing an improvement to full and fractional factorial designs.

Two steps are needed to apply that method:

- A step of generation of the Taguchi design during which it is possible to choose from a list of designs according to the number of factors and the levels of these.
- Once the Taguchi design is obtained, you can analyze it to identify the control factor parameters that minimize the variation of the response. This second step is achievable with the tool Analysis of a Taguchi design.

Taguchi's method

G. Taguchi's approach is one of the most widely used engineering techniques in the world. This method was aimed at developing products that functioned well despite natural variations in materials, operators and environment.

It consists of using orthogonal tables which have been pre-established by G. Taguchi and depend on the number of runs to be performed, the number of factors in the model, and the number of levels by factors.

The Taguchi method divides optimization problems into two categories: the static method and the dynamic method.

The purpose of static Taguchi designs is to determine the best levels of control factors so that the output is at the desired value.

Dynamic designs have a signal factor. The objective is to determine the best levels of control factors in order to improve the relationship between this signal factor and an output response.

Depending on the number of factors and their levels, XLSTAT offers, in a new dialog box, a list of designs that can be made. These designs have a particular notation defined like that: L(nbRuns)(nbLevel^nbFactor)

Where nbRuns = number of runs, nbLevel = number of levels for each factor, and nbFactors = number of factors.

If the notation is of this form: L(nbRuns)(nbLevel^nbFactors nbLevel^nbFactors), the design contains mixed-levels factors.

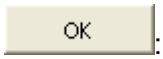
For example, a L9(3^3) design is a design having 9 runs and 3 factors at 3 levels. A L18(3^3 6^1) design is a design having 18 runs, 3 factors with 3 levels and 1 factor with 6 levels.

The list of designs available in XLSTAT is as follows:

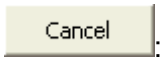
- $L4(2^3)$
- $L8(2^7)$
- $L9(3^4)$
- $L12(2^{11})$
- $L16(2^{15})$
- $L16(4^5)$
- $L25(5^6)$
- $L27(3^{13})$
- $L32(2^{31})$
- $L32(2^1 4^9)$
- $L36(2^{11} 3^{12})$
- $L36(2^3 3^{13})$
- $L50(2^1 4^{11})$
- $L8(4^4 2^1)$
- $L16(4^1 2^{12})$
- $L16(4^2 2^9)$
- $L16(4^3 2^6)$
- $L16(4^4 2^3)$
- $L18(2^1 3^7)$
- $L18(3^6 6^1)$

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Factors/Categories Table: Enter the table listing the factors and their categories.

Number of responses: Enter the number of responses from your analysis.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selections contains a label.

Onglet **Options** :

Add signal factor: Check this option to add a signal factor to the Taguchi design.

- **Number of levels**: Select the number of levels of the signal factor.
- **Signal labels**: Select the labels for each of the signal factor levels. These labels must be numeric.

Interactions: Check this option to add interactions to the Taguchi design.

Outputs tab:

Coded design: Enable this option if you want to display the coded design, ie the Taguchi design containing 1, 2, 3, ... rather than the categories of your factors.

Results

Variable Information: This table summarizes all information about the selected factors.

Design of experiment: This table shows the Taguchi design. Empty columns are used to fill in the responses.

Start the analysis: Once the responses columns are filled, you can click on the "Run the analysis" button to open the prefilled dialog box allowing to perform the analysis of the Taguchi design.

Coded design: This table displays Taguchi's coded design, containing 1, 2, 3, ... rather than the categories of your factors.

Example

An example of a Taguchi design analysis is available on XLSTAT Help Center:

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

References

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.

Analysis of a Taguchi design

Use this tool to analyze a Taguchi design to find products that are robust and insensitive to the natural variability of the environment and processes.

In this section:

Description

The Taguchi method is a method introduced by Genichi Taguchi (Genichi and Wu, 1980) which is a method of experimental design providing an improvement to full and fractional factorial designs.

It consists in using orthogonal tables, which have been pre-established by G. Taguchi, which depend on the number of runs to be performed, the number of factors in the model, and the number of levels by factors.

In conventional experimental designs, the goal is to identify the factors that affect the average response and control them at the desired levels. Taguchi's experimental designs deal with the mean and variability of measured characteristic values through the use of signal to noise ratios (S/N).

Taguchi's method

G. Taguchi's approach is one of the most widely used engineering techniques in the world. This method was aimed at developing products that functioned well despite natural variations in materials, operators and environment.

In order to find products that are robust and insensitive to variability, XLSTAT offers to study three parameters: the signal to noise ratios, the means and the standard deviations.

XLSTAT also allows you to adjust the linear model for these three parameters. A table of estimated regression coefficients will then be displayed, and it will be possible to determine which factors have statistically significant values at $\alpha = 0.05$ threshold.

Signal to Noise ratios

The signal to noise ratio is different depending on whether a static Taguchi design or a dynamic design is used.

Static signal to noise ratio :

Traditionally, only one output of a product or process is used in research and development. In this case, a non-dynamic signal to noise ratio is used to improve the robustness of the product in question. For a non-dynamic signal to noise ratio, there are two problems: reducing variability, and adjusting the average.

Depending on the purpose of your experiment, several signal to noise ratios are available in XLSTAT:

- **Larger is better**: Select this option if the intention is to maximize the answer and there are no negative data. To calculate this ratio we use the following formula:

$$S/N = -10 \times \log\left(\frac{\sum(\frac{1}{Y^2})}{n}\right)$$

where Y is the answer for the given factor level combination and n is the number of responses in the factor level combination.

- **Nominal is best: type I**: select this option if there is no negative data. For a "Nominal is best" application, optimization is carried out in two stages. The first is to maximize the signal to noise ratio, the second to adjust the average. To calculate this ratio we use the following formula:

$$S/N = 10 \times \log\left(\frac{\bar{Y}^2}{s^2}\right)$$

where $\bar{Y}$ is the average of the responses for the given factor level combination, s is the standard deviation of the responses for the given factor combination and n is the number of responses in the combination of factor levels.

- **Nominal is best: type II**: Select this option when the results of an experiment include negative values. Thus, the positive and negative values cancel each other out. Sensitivity or average information can not be obtained. We then choose this type of signal to noise ratio, which only shows the variability, however, it is less informative than the type I. To calculate this ratio we use the following formula:

$$S/N = -10 \times \log(s^2)$$

where s is the standard deviation of the responses for all the noise factors of the given factor level combination.

- **Smaller is better**: Select this option if the intention is to minimize the response and there is no negative data. To calculate this ratio we use the following formula:

$$S/N = -10 \times \log\left(\frac{\sum Y^2}{n}\right)$$

where Y is the answer for the given factor level combination and n is the number of responses in the factor level combination.

Dynamic signal to noise ratio :

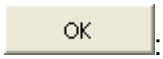
In the case of a dynamic Taguchi design, XLSTAT uses a zero-point proportional equation. That is, the input is equal to zero, and the output must pass through the origin. To calculate this ratio we use the following formula:

$$S/B = 10 \times \log\left(\frac{slope^2}{MSE}\right)$$

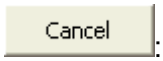
where the slope is the rate of change of the response relative to the signal factor, and MSE is the Mean Square Error.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Responses: Select the answers for the different levels of the noise factors.

Taguchi design: Select the Taguchi design columns corresponding to the different control factors.

Signal factor: Check this option if you selected a signal factor when creating the Taguchi design. Then select the column corresponding to the signal factor.

Range: Check this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Check this option to display the results in a new worksheet in the active workbook.

Workbook: Check this option to display the results in a new workbook.

Variable labels: Check this option if the first line of the selections contains a label.

Options tab:

Signal to Noise ratio: Enable the type of signal to noise ratios you want to calculate from the following 4:

- **Larger is better**
- **Nominal is best: Type II**
- **Nominal is best: Type I**
- **Smaller is better**

Adjust Linear Model: Enable this option if you want to adjust the linear model for the coefficients selected in the **Outputs** tab.

- **Interactions**: Check this option if you want to add interactions to the model. In order to have consistent results, it is necessary to have selected the interactions when creating the design.

Outputs tab:

Signal to Noise ratio: Enable this option if you want to display the response table for signal to noise ratios.

Means: Enable this option if you want to display the response table for the means. This option is only available in the case of a static Taguchi design.

Slope: Enable this option if you want to display the response table for slopes. This option is only available in the case of a dynamic Taguchi design.

Standard Deviations: Enable this option if you want to display the response table for standard deviations.

Charts tab:

Signal to Noise ratio: Enable this option if you want to display the main effects chart for signal to noise ratios.

Averages: Enable this option if you want to display the main effects chart for the means. This option is only available in the case of a static Taguchi design.

Slope: Enable this option if you want to display the main effects chart for slopes. This option is only available in the case of a dynamic Taguchi design.

Standard Deviations: Enable this option if you want to display the main effects chart for standard deviations.

Results

Variable Information: This table summarizes all information about the selected attributes.

Signal to Noise ratio: This table displays the responses for signal to noise ratios.

Means: This table displays the responses for means.

Slopes: This table displays the responses for the slopes.

Standard Deviations: This table displays the responses for standard deviations.

Regression of variable:

Goodness of fit statistics: The statistics relating to the fitting of the regression model are shown in this table:

- **Observations**: The number of observations used in the calculations. In the formulas shown below, n is the number of observations.
- **Sum of weights**: The sum of the weights of the observations used in the calculations. In the formulas shown below, W is the sum of the weights.
- **DF**: The number of degrees of freedom for the chosen model (corresponding to the error part).
- **R²**: The determination coefficient for the model. This coefficient, whose value is between 0 and 1, is only displayed if the constant of the model has not been fixed by the user. Its value is defined by:

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ with } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

The R^2 is interpreted as the proportion of the variability of the dependent variable explained by the model. The nearer R^2 is to 1, the better is the model. The problem with the R^2 is that it does not take into account the number of variables used to fit the model.

- **Adjusted R²**: The adjusted determination coefficient for the model. The adjusted R^2 can be negative if the R^2 is near to zero. Its value is defined by:

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}$$

The adjusted R² is a correction to the R² which takes into account the number of variables used in the model.

- **MSE**: The mean squared error (MSE) is defined by:

$$MSE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

- **RMSE**: The root mean square of the errors (RMSE) is the square root of the MSE.
- **MAPE**: The *Mean Absolute Percentage Error* is calculated as follows:

$$MAPE = \frac{100}{W} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- **DW**: The Durbin-Watson statistic is defined by:

$$DW = \frac{\sum_{i=2}^{n} [(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}$$

This coefficient is the order 1 autocorrelation coefficient and is used to check that the residuals of the model are not autocorrelated, given that the independence of the residuals is one of the basic hypotheses of linear regression. The user can refer to a table of Durbin-Watson statistics to check if the independence hypothesis for the residuals is acceptable.

- **Cp**: Mallows Cp coefficient is defined by:

$$Cp = \frac{SSE}{\hat{\sigma}} + 2p^* - W$$

where SSE is the sum of the squares of the errors for the model with p explanatory variables and $\hat{\sigma}$ is the estimator of the variance of the residuals for the model comprising all the explanatory variables. The nearer the Cp coefficient is to $p*$, the less the model is biased.

- **AIC**: Akaike's Information Criterion is defined by:

$$AIC = W \ln(\frac{SSE}{W}) + 2p^*$$

This criterion, proposed by Akaike (1973) is derived from the information theory and uses Kullback and Leibler's measurement (1951). It is a model selection criterion which penalizes models for which adding new explanatory variables does not supply sufficient information to the model, the information being measured through the MSE. The aim is to minimize the AIC criterion.

- **SBC**: Schwarz's Bayesian Criterion is defined by:

$$SBC = W \ln(\frac{SSE}{W}) + \ln(W)p^*$$

This criterion, proposed by Schwarz (1978) is similar to the AIC, and the aim is to minimize it.

- **PC**: Amemiya's Prediction Criterion is defined by:

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}$$

This criterion, proposed by Amemiya (1980) is used, like the adjusted R^2 to take account of the parsimony of the model.

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. The explanatory power is evaluated by comparing the fit (as regards least squares) of the final model with the fit of the rudimentary model including only a constant equal to the mean of the dependent variable.

The table of **Type I SS** values is used to visualize the influence that progressively adding explanatory variables has on the fitting of the model, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. The sums of squares in the Type I table always add up to the model SS. Note: the order in which the variables are selected in the model influences the values obtained.

The table of **Type II SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained.

The table of **Type III SS** values is used to visualize the influence that removing an explanatory variable has on the fitting of the model, all other variables being retained, expect those were the effect is present (interactions), as regards the sum of the squares of the errors (SSE), the mean squared error (MSE), Fisher's F, or the probability associated with Fisher's F. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. Note: unlike Type I SS, the order in which the variables are selected in the model has no influence on the values obtained. Type II and Type III are identical if there are no interactions or if the design is balanced.

The **parameters of the model** table displays the estimate of the parameters, the corresponding standard error, the Student’s t, the corresponding probability, as well as the confidence interval

The **equation of the model** is then displayed to make it easier to read or re-use the model.

Example

An example of a Taguchi design analysis is available on XLSTAT Help Center:

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

References

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.

Survival analysis

Kaplan-Meier analysis

Use this tool to build a population survival curve, and to obtain essential statistics such as the median survival time. Kaplan-Meier analysis, which main result is the Kaplan-Meier table, is based on irregular time intervals, contrary to the Life table analysis, where the time intervals are regular.

In this section:

Description

The Kaplan Meier method (also called product-limit) analysis belongs to the descriptive methods of survival analysis, as does Life table analysis. The life table analysis method was developed first, but the Kaplan-Meier method has been shown to be superior in many cases.

Kaplan-Meier analysis allows to quickly obtain a population survival curve and essential statistics such as the median survival time. Kaplan-Meier analysis, which main result is the Kaplan-Meier table, is based on irregular time intervals, contrary to the Life table analysis, where the time intervals are regular.

Kaplan-Meier analysis is used to analyze how a given population evolves with time. This technique is mostly applied to survival data and product quality data. There are three main reasons why a population of individuals or products may evolve: some individuals die (products fail), some other go out of the surveyed population because they get healed (repaired) or because their trace is lost (individuals move from location, the study is terminated, …). The first type of data is usually called "failure data", or "event data", while the second is called "censored data".

There are several types of censoring of survival data:

- Left censoring: when an event is reported at time t=t(i), we know that the event occurred at t < t(i).
- Right censoring: when an event is reported at time t=t(i), we know that the event occurred at t > t(i), if it ever occurred.

- Interval censoring: when an event is reported at time t=t(i), we know that the event occurred during [t(i-1); t(i)].
- Exact censoring: when an event is reported at time t=t(i), we know that the event occurred exactly at t=t(i).

The Kaplan Meier method requires that the observations are independent. Second, the censoring must be independent: if you consider two random individuals in the study at time t-1, if one of the individuals is censored at time t, and if the other survives, then both must have equal chances to survive at time t. There are four different types of independent censoring:

- Simple type I: all individuals are censored at the same time or equivalently individuals are followed during a fixed time interval.
- Progressive type I: all individuals are censored at the same date (for example, when the study terminates).
- Type II: the study is continued until n events have been recorded.
- Random: the time when a censoring occurs is independent of the survival time.

The Kaplan Meier analysis allows to compare populations, through their survival curves. For example, it can be of interest to compare the survival times of two samples of the same product produced in two different locations. Tests can be performed to check if the survival curves have arisen from identical survival functions. These results can later be used to model the survival curves and to predict probabilities of failure.

Confidence interval

Computing confidence intervals for the survival function can be done using three different methods :

Greenwood's method:

$$S(T) \pm z_{1-\alpha/2}\sqrt{\frac{\mathrm{Var}(S(T))}{S^2(T)}}$$

Exponential Greenwood's method:

$$\exp(-\exp(\log(-\log(S(T))) \pm z_{1-\alpha/2}\sqrt{\mathrm{var}(S(T))}))$$

Log-transformed method: $S(T)^{1/\theta}, S(T)^{\theta}$ with

$$\theta = exp\left(\frac{z_{1-\alpha/2}\sqrt{\frac{\mathrm{Var}(S(T))}{S^2(T)}}}{log(S(T))}\right)$$

These three approaches give similar results, but the last ones will be preferred when samples are small.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

General tab:

Date data: Select the data that correspond to the times or the dates when the events or the censoring are recorded. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Weighted data: Activate this option if for a given time, several events are recorded on the same row (for example, at time t=218, 10 failures and 2 censured data have been observed). If you activate this option, the "Event indicator" field replaces the "Status variable" field, and the "Censoring indicator" field replaces the "Event code" and "Censured code" boxes.

Status indicator: Select the data that correspond to an event or censoring data. This field is not available if the "Weighted data" option is checked. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Event code: Enter the code used to identify an event data within the Status variable. Default value is 1.

Censored code Event code: Enter the code used to identify an event data within the Status variable. Default value is 1.

Censored code: Enter the code used to identify a censored data within the Status variable. Default value is 0.

Event indicator: Select the data that correspond to the counts of events recorded at each time. Note: this option is available only if the "weighted data" option is selected. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Censoring indicator: Select the data that correspond to the counts of right-censored data recorded at a given time. Note: this option is available only if the "weighted data" option is selected. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if the row and column labels have been selected.

Options tab:

Significance level (%): Enter the significance level for the comparison tests (default value 5%). This value is also used to determine the confidence intervals around the estimated statistics.

Confidence interval: Choose the method to use to compute the confidence interval to be displayed in the outputted table.

Data options tab:

Missing data:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Groups:

By group analysis: Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis on each group separately.

- **Compare**: Activate this option if want to compare the survival curves, and perform the comparison tests.

Filter: Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis for some groups that you will be able to select in a separate dialog box during the computations. If the "By group analysis" option is also activated, XLSTAT will perform the analysis for each group separately, only for the selected subset of groups.

Charts tab:

Survival distribution function: Activate this option to display the charts corresponding to the survival distribution function.

-Log(SDF): Activate this option to display the –Log() of the survival distribution function (SDF).

Log(-Log(SDF)): Activate this option to display the Log(–Log()) of the survival distribution function.

1 - Survival distribution function: Activate this option to display the charts corresponding to 1 - survival distribution function.

Censored data: Activate this option to identify on the charts the times when censored data have been recorded (the identifier can be a hollowed circle "o" or a "+").

Results

Basic statistics: This table displays the total number of observations, the number of events, and the number of censored data.

Kaplan-Meier table: This table displays the various results obtained from the analysis, including:

- **Interval start lime**: lower bound of the time interval.
- **At risk**: number of individuals that were at risk.
- **Events**: number of events recorded.
- **Censored**: number of censored data recorded.
- **Proportion failed**: proportion of individuals who "failed" (the event did occur).
- **Survival rate**: proportion of individuals who "survived" (the event did not occur).
- **Survival distribution function (SDF)**: Probability of an individual to survive until at least the time of interest. Also called cumulative survival distribution function, or survival curve.
- **Survival distribution function standard error:** standard error of the previous
- **Survival distribution function confidence interval :** confidence interval of the previous.

Mean and Median residual lifetime: A first table displays the mean residual lifetime, the standard error, and a confidence range. A second table displays statistics (estimator, and confidence range) for the 3 quartiles including the median residual lifetime (50%). The median residual lifetime is one of the key results of the Kaplan-Meier analysis as it allows to evaluate the time remaining for half of the population to "fail".

Charts: Depending on the selected options, up to four charts are displayed: Survival distribution function (SDF), -Log(SDF), Log(-Log(SDF)) and 1 - SDF.

If the "Compare" option has been activated in the dialog box, XLSTAT displays the following results:

Test of equality of the survival functions: This table displays the statistics for three different tests: the Log-rank test, the Wilcoxon test, and the Tarone Ware test. These tests are based on a Chi-square test. The lower the corresponding p-value, the more significant the differences between the groups.

If the pvalue obtained by the log-rank test is significant at alpha = 5% threshold, then multiple pairwise comparison tests are performed on groups. We perform a Dunn-Sidak test which is a derivative of Bonferroni test and is more efficient in certain situations.

$$\alpha' = 1 - (1 - \alpha)^{1/g}.$$

where g is the number of groups.

Charts: Depending on the selected options, up to four charts with one curve for each group are displayed: Survival distribution function (SDF), -Log(SDF), Log(-Log(SDF)), 1-SDF.

Example

An example of survival analysis based on the Kaplan-Meier method is available on the XLSTAT Help Center:

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

References

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.

Life tables

Use this tool to build a survival curve for a given population, and to obtain essential statistics such as the median survival time. Life table analysis, which main result is the life table (also named actuarial table), works on regular time intervals, contrary to the Kaplan Meier analysis, where the time intervals are taken as they are in the data set. XLSTAT enables you to take into account censored data, and grouping information.

In this section:

Description

Life table analysis belongs to the descriptive methods of survival analysis, as well as Kaplan Meier analysis. The life table analysis method was developed first, but the Kaplan-Meier method has been shown to be superior in many cases.

Life table analysis allows to quickly obtain a population survival curve and essential statistics such as the median survival time. Life table analysis, which main result is the life table (also called actuarial table) works on regular time intervals, contrary to the Kaplan Meier analysis, where the time intervals are taken as they are in the data set.

Life table analysis allows to analyze how a given population evolves with time. This technique is mostly applied to survival data and product quality data. There are three main reasons why a population of individuals or products may evolve: some individuals die (products fail), some other go out of the surveyed population because they get healed (repaired) or because their trace is lost (individuals move from location, the study is terminated, …). The first type of data is usually called "failure data", or "event data", while the second is called "censored data".

There are several types of censoring of survival data:

Left censoring: when an event is reported at time t=t(i), we know that the event occurred at t < t(i).

Right censoring: when an event is reported at time t=t(i), we know that the event occurred at t > t(i), if it ever occurred.

Interval censoring: when an event is reported at time t=t(i), we know that the event occurred during [t(i-1); t(i)].

Exact censoring: when an event is reported at time t=t(i), we know that the event occurred exactly at t=t(i).

The life table method requires that the observations are independent. Second, the censoring must be independent: if you consider two random individuals in the study at time t-1, if one of the individuals is censored at time t, and if the other survives, then both must have equal chances to survive at time t. There are four different types of independent censoring:

Simple type I: all individuals are censored at the same time or equivalently individuals are followed during a fixed time interval.

Progressive type I: all individuals are censored at the same date (for example, when the study terminates).

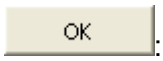
Type II: the study is continued until n events have been recorded.

Random: the time when a censoring occurs is independent of the survival time.

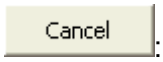
The life table method allows to compare populations, through their survival curves. For example, it can be of interest to compare the survival times of two samples of the same product produced in two different locations. Tests can be performed to check if the survival curves have arisen from identical survival functions. These results can later be used to model the survival curves and to predict probabilities of failure.

Dialog box

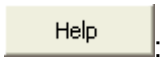
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Date data: Select the data that correspond to the times or the dates when the events or the censoring are recorded. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Weighted data: Activate this option if for a given time, several events are recorded on the same row (for example, at time t=218, 10 failures and 2 censured data have been observed). If you activate this option, the "Event indicator" field replaces the "Status variable" field, and the "Censoring indicator" field replaces the "Event code" and "Censured code" boxes.

Status indicator: Select the data that correspond to an event or censoring data. This field is not available if the "Weighted data" option is checked. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Event code: Enter the code used to identify an event data within the Status variable. Default value is 1.

Censored code Event code: Enter the code used to identify an event data within the Status variable. Default value is 1.

Censored code: Enter the code used to identify a censored data within the Status variable. Default value is 0.

Event indicator: Select the data that correspond to the counts of events recorded at each time. Note: this option is available only if the "weighted data" option is selected. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Censoring indicator: Select the data that correspond to the counts of right-censored data recorded at a given time. Note: this option is available only if the "weighted data" option is selected. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if the row and column labels have been selected.

Options tab:

Significance level (%): Enter the significance level for the comparison tests (default value 5%). This value is also used to determine the confidence intervals around the estimated statistics.

Time intervals:

- **Constant width**: Activate this option if want to enter the constant interval width. In this case, the lower bound is automatically set to 0.

- **User defined**: Activate this option to define the intervals that should be used to perform the life table analysis. Then select the data that correspond to the lower bound of the first interval and to the upper bounds of all the intervals.

Data options tab:

Missing data:

- **Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.
- **Remove observations**: Activate this option to remove the observations with missing data.

Groups:

By group analysis: Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis on each group separately.

- **Compare**: Activate this option if want to compare the survival curves, and perform the comparison tests.

Filter: Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis for some groups that you will be able to select in a separate dialog box during the computations. If the "By group analysis" option is also activated, XLSTAT will perform the analysis for each group separately, only for the selected subset of groups.

Charts tab:

Survival distribution function: Activate this option to display the charts corresponding to the survival distribution function.

-Log(SDF): Activate this option to display the –Log() of the survival distribution function (SDF).

Log(-Log(SDF)): Activate this option to display the Log(–Log()) of the survival distribution function.

Censored data: Activate this option to identify on the charts the times when censored data have been recorded (the identifier can be a hollowed circle "o" or a "+").

Results

Basic statistics: This table displays the total number of observations, the number of events, and the number of censored data.

Life table: This table displays the various results obtained from the analysis, including:

- **Interval**: Time interval.

- **At risk**: Number of individuals that were at risk during the time interval.
- **Events**: Number of events recorded during the time interval.
- **Censored**: Number of censored data recorded during the time interval.
- **Effective at risk**: Number of individuals that were at risk at the beginning of the interval minus half of the individuals who have been censored during the time interval.
- **Survival rate**: Proportion of individuals who "survived" (the event did not occur) during the time interval. Ratio of individuals who survived over the individuals who were "effective at risk".
- **Conditional probability of failure**: Ratio of individuals who failed over the individuals who were "effective at risk".
- Standard error of the conditional probability: Standard error of the previous.
- **Survival distribution function (SDF)**: Probability of an individual to survive until at least the time interval of interest. Also called survivor function.
- Standard error of the survival function: standard error of the previous.
- Probability density function: estimated density function at the midpoint of the interval.
- Standard error of the probability density: standard error of the previous.
- **Hazard rate**: estimated hazard rate function at the midpoint of the interval. Also called failure rate. Corresponds to the failure rate for the survivors.
- **Standard error of the hazard rate**: Standard error of the previous.
- **Median residual lifetime**: Amount of time remaining to reduce the surviving population (individuals at risk) by one half. Also called median future lifetime.
- Median residual lifetime standard error: Standard error of the previous.

Median residual lifetime: Table displaying the median residual lifetime at the beginning of the experiment, and its standard error. This statistic is one of the key results of the life table analysis as it allows to evaluate the time remaining for half of the population to "fail".

Charts: Depending on the selected options, up to five charts are displayed: Survival distribution function (SDF), Probability density function, Hazard rate function, -Log(SDF), Log(-Log(SDF)).

If the "Compare" option has been activated in the dialog box, XLSTAT displays the following results:

Test of equality of the survival functions: This table displays the statistics for three different tests: the Log-rank test, the Wilcoxon test, and the Tarone Ware test. These tests are based on

a Chi-square test. The lower the corresponding p-value, the more significant the differences between the groups.

If the pvalue obtained by the log-rank test is significant at alpha = 5% threshold, then multiple pairwise comparison tests are performed on groups. We perform a Dunn-Sidak test which is a derivative of Bonferroni test and is more efficient in certain situations.

$$\alpha' = 1 - (1 - \alpha)^{1/g}.$$

where g is the number of groups.

Charts: Depending on the selected options, up to five charts with one curve for each group are displayed: Survival distribution function (SDF), Probability density function, Hazard rate function, -Log(SDF), Log(-Log(SDF)).

Example

An example of survival analysis by the mean of life tables is available on the XLSTAT Help Center:

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

References

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.

Nelson-Aalen analysis

Use this tool to build cumulative hazard curves using the Nelson-Aalen method. The Nelson-Aalen method allows to estimate the hazard functions based on irregular time intervals, contrary to the Life table analysis, where the time intervals are regular.

In this section:

Description

The Nelson-Aalen analysis method belongs to the descriptive methods for survival analysis. With the Nelson-Aalen approach you can quickly obtain a curve of cumulative hazard. The Nelson-Aalen method enables to estimate the hazard functions based on irregular time intervals.

Nelson-Aalen analysis is used to analyze how a given population evolves with time. This technique is mostly applied to survival data and product quality data. There are three main reasons why a population of individuals or products may evolve: some individuals die (products fail), some other go out of the surveyed population because they get healed (repaired) or because their trace is lost (individuals move from location, the study is terminated, …). The first type of data is usually called "failure data", or "event data", while the second is called "censored data".

There are several types of censoring of survival data:

- **Left censoring**: when an event is reported at time$t = t(i)$, we know that the event occurred at $t < t(i)$.
- **Right censoring**: when an event is reported at time $t = t(i)$, we know that the event occurred at $t > t(i)$, if it ever occurred.
- **Interval censoring**: when an event is reported at time $t = t(i)$, we know that the event occurred during $[t(i-1); t(i)]$.
- **Exact censoring**: when an event is reported at time $t = t(i)$, we know that the event occurred exactly at $t = t(i)$.

The Nelson-Aalen method requires that the observations are independent. Second, the censoring must be independent: if you consider two random individuals in the study at time $t-1$, if one of the individuals is censored at time t, and if the other survives, then both must have equal chances to survive at time t. There are four different types of independent censoring:

- **Simple type I**: all individuals are censored at the same time or equivalently individuals are followed during a fixed time interval.
- **Progressive type I**: all individuals are censored at the same date (for example, when the study terminates).
- **Type II**: the study is continued until n events have been recorded.
- **Random**: the time when a censoring occurs is independent of the survival time.

The Nelson-Aalen analysis allows to compare populations, through their hazards curves. Nelson-Aalen estimator should be preferred to Kaplan-Meier estimator when analyzing cumulative hazard functions. When analyzing cumulative survival functions, Kaplan-Meier estimator should be preferred.

The cumulative hazard function is: $H(T) = \sum_{T_i \leq T} \frac{d_i}{r_i}$

with d_i being the number of observation falling at time T_i and r_i, the number of observation at risk (still in the study) at time T_i.

Several different variance estimators are available:

- Simple: $\mathrm{var}(H(T)) = \sum_{T_i \leq T} \frac{d_i}{r_i^2}$
- Plug-in: $\mathrm{var}(H(T)) = \sum_{T_i \leq T} \frac{d_i(r_i - d_i)}{r_i^3}$
- Binomial: $\mathrm{var}(H(T)) = \sum_{T_i \leq T} \frac{d_i(r_i - d_i)}{r_i^2(r_i - 1)}$

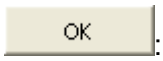
Confidence intervals can also be obtained :

- Greenwood's method: $H(T) \pm z_{1-\alpha/2}\sqrt{\mathrm{Var}(H(T))}$
- Log-transformed method: $H(T)/\phi,\ H(T) \times \phi$ with $\phi = \exp(\frac{z_{1-\alpha/2}\sqrt{\mathrm{var}(H(T))}}{H(T)})$

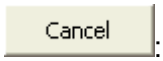
The second one will be preferred with small samples.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Date data: Select the data that correspond to the times or the dates when the events or the censoring are recorded. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Weighted data: Activate this option if for a given time, several events are recorded on the same row (for example, at time t=218, 10 failures and 2 censured data have been observed). If you activate this option, the "Event indicator" field replaces the "Status variable" field, and the "Censoring indicator" field replaces the "Event code" and "Censured code" boxes.

Status indicator: Select the data that correspond to an event or censoring data. This field is not available if the "Weighted data" option is checked. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Event code: Enter the code used to identify an event data within the Status variable. Default value is 1.

Censored code Event code: Enter the code used to identify an event data within the Status variable. Default value is 1.

Censored code: Enter the code used to identify a censored data within the Status variable. Default value is 0.

Event indicator: Select the data that correspond to the counts of events recorded at each time. Note: this option is available only if the "weighted data" option is selected. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Censoring indicator: Select the data that correspond to the counts of right-censored data recorded at a given time. Note: this option is available only if the "weighted data" option is selected. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if the column labels have been selected.

Options tab:

Significance level (%): Enter the significance level for the comparison tests (default value 5%). This value is also used to determine the confidence intervals around the estimated statistics.

Variance: Choose the method to use to compute the variance to be displayed in the outputted table.

Confidence interval: Choose the method to use to compute the confidence interval to be displayed in the outputted table.

Data options tab:

Missing data:

- **Do not accept missing data**: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.
- **Remove observations**: Activate this option to remove the observations with missing data.
- **Ignore missing data**: Activate this option to ignore missing data.

Groups:

By group analysis: Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis on each group separately.

- **Compare**: Activate this option if want to compare the survival curves, and perform the comparison tests.

Filter: Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis for some groups that you will be able to select in a separate dialog box during the computations. If the "By group analysis" option is also activated, XLSTAT will perform the analysis for each group separately, only for the selected subset of groups.

Charts tab:

Cumulative hazard function: Activate this option to display the charts corresponding to the cumulative hazard function.

Survival distribution function: Activate this option to display the charts corresponding to the survival distribution function.

Log(Cumulative hazard function): Activate this option to display the Log() of the cumulative hazard function.

Censored data: Activate this option to identify on the charts the times when censored data have been recorded (the identifier can be a hollowed circle "o" or a "+").

Results

Basic statistics: This table displays the total number of observations, the number of events, and the number of censored data.

Nelson-Aalen table: This table displays the various results obtained from the analysis, including:

- **Interval start lime**: lower bound of the time interval.
- **At risk**: number of individuals that were at risk.
- **Events**: number of events recorded.
- **Censored**: number of censored data recorded.
- **Cumulative hazard function**: hazard associated with an individual at the considered time.
- **Cumulative hazard function error:** standard error of the previous
- **Cumulative hazard function confidence interval:** confidence interval of the previous
- **Survival distribution function:** probability for an individual to survive until the considered time (calculated as $S(T) = \exp(-H(T))$).

Charts: Depending on the selected options, up to three charts are displayed: Cumulative hazard function, survival distribution function, and Log(Hazard function).

If the "Compare" option has been activated in the dialog box, XLSTAT displays the following results:

Test of equality of the survival functions: This table displays the statistics for three different tests: the Log-rank test, the Wilcoxon test, and the Tarone Ware test. These tests are based on a Chi-square test. The lower the corresponding p-value, the more significant the differences between the groups.

If the pvalue obtained by the log-rank test is significant at alpha = 5% threshold, then multiple pairwise comparison tests are performed on groups. We perform a Dunn-Sidak test which is a derivative of Bonferroni test and is more efficient in certain situations.

$$\alpha' = 1 - (1 - \alpha)^{1/g}.$$

where g is the number of groups.

Charts: Depending on the selected options, up to three charts with one curve for each group are displayed: Cumulative hazard function, survival distribution function, and Log(Hazard function).

Example

An example of survival analysis based on the Nelson-Aalen method is available on the XLSTAT Help Center:

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

References

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.

Cumulative incidence

Use this tool to analyze survival data when competing risks are present. The cumulative incidence allows to estimate the impact of an event when several competitive events may occur. The time intervals should not necessarily be regular. XLSTAT allows the treatment of censored data with competing risks and to compare different groups within the population.

In this section:

Description

The cumulative incidence allows estimating the impact when several competitive events may occur. It is usually called competing risks case. The time intervals should not necessarily be regular. XLSTAT allows the treatment of censored data in competing risks and to compare different groups within the population.

For a given period, the cumulative incidence is the probability that an observation still included in the analysis at the beginning of this period will be affected by an event during the period. It is especially appropriate in the case of competing risks, that is to say, when several types of events may occur.

This technique is used for the analysis of survival data, whether individuals (cancer research, for example) or products (resistance time of a production tool, for example): some individuals die (in this case we will have 2 causes of death: from the disease or an other cause), the products break (in this case we can model different breaking points), but others leave the study because they heal, you lose track of them (moving for example) or because the study was discontinued. The first type of data is usually called "failure data", or "event data", while the second is called "censored data".

There are several types of censoring of survival data:

- Left censoring: when an event is reported at time $t = t(i)$, we know that the event occurred at $t \times t(i)$.
- Right censoring: when an event is reported at time $t = t(i)$, we know that the event occurred at $t \times t(i)$, if it ever occurred.
- Interval censoring: when an event is reported at time $t = t(i)$, we know that the event occurred during $[t(i-1); t(i)]$.

- Exact censoring: when an event is reported at time $t = t(i)$, we know that the event occurred exactly at $t = t(i)$.

The cumulative incidence method requires that the observations are independent. Second, the censoring must be independent: if you consider two random individuals in the study at time $t - 1$, if one of the individuals is censored at time t, and if the other survives, then both must have equal chances to survive at time t. There are four different types of independent censoring:

- Simple type I: all individuals are censored at the same time or equivalently individuals are followed during a fixed time interval.
- Progressive type I: all individuals are censored at the same date (for example, when the study terminates).
- Type II: the study is continued until n events have been recorded.
- Random: the time when a censoring occurs is independent of the survival time.

When working with competing risks, the different types of events can happen only once, after the event has occurred, the observation is withdrawn from the analysis. We can calculate the risk of occurrence of an event in the presence of competitive events. XLSTAT allows you to compare the types of events but also to take account of groups of observations (depending on the treatment administered, for example).

The cumulative incidence function is: $I_k(T) = \sum_{T_j \leq T} \hat{S}(T_{j-1}) \frac{d_{kj}}{n_j}$ for event k at time T. With $\hat{S}(T_{j-1})$ being the survival distribution function obtained using the Kaplan-Meier estimator, d_{kj} being the number of observation failing with event k at time T_j and n_j, the number of observation at risk (still in the study) at time T_j.

Variance estimator is:

$$\begin{aligned} Var(I_k(T)) = & \sum_{T_j \leq T} \left[\left(I_k(T) - I_k(T_j)\right)^2 \frac{d_j}{n_j(n_j - d_j)} \right] \\ & + \sum_{T_j \leq T} \left[\left(\hat{S}(T_{j-1})\right)^2 \frac{(n_j - d_j)}{n_j} \frac{d_{kj}}{n_j^2} \right] \\ & - 2 \sum_{T_j \leq T} \left[\left(I_k(T) - I_k(T_j)\right) \hat{S}(T_{j-1}) \frac{d_j}{n_j^2} \right] \end{aligned}$$

Confidence intervals are obtained using:

$$I_k(T)^{exp\left(\frac{\pm z_{\alpha/2}\sqrt{Var(I_k(T))}}{I_k(T) log(I_k(T))}\right)}$$

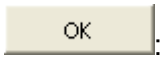
Gray test for group comparison

Gray test is used to compare groups in a cumulative incidence framework. When competing risks are present, a classic comparison of groups test cannot be applied. Gray developed a test for that case. It is based on a k-sample test that compares the cumulative incidence of a particular type of failure among different groups. For a complete presentation of that test, see Gray (1988).

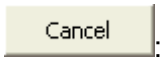
A p-value for each failure type is obtained for all the groups being studied.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Date data: Select the data that correspond to the times or the dates when the events or the censoring are recorded. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Status indicator: Select the data that correspond to an event or censoring data. This field is not available if the "Weighted data" option is checked. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Censored code: Enter the code used to identify a censored data within the Status variable. Default value is 0.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if the column labels have been selected.

Groups: Activate this option if you want to group the data. Then select the data that correspond to the group to which each observation belongs.

Gray test: Activate this option if you want to perform a Gray test to compare cumulative incidence associated to groups of observations for each failure type.

Options tab:

Significance level (%): Enter the significance level for the comparison tests (default value 5%). This value is also used to determine the confidence intervals around the estimated statistics.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Charts tab:

Cumulative incidence function: Activate this option to display the charts corresponding to the cumulative incidence function.

Survival distribution function: Activate this option to display the charts corresponding to the survival distribution function.

Censored data: Activate this option to identify on the charts the times when censored data have been recorded (the identifier is a hollowed circle "o").

Results

Basic statistics: This table displays the total number of observations, the number of events, and the number of censored data.

Each table and plots are displayed for each event type.

Cumulative incidence: This table displays the various results obtained from the analysis, including:

- **Interval start lime**: lower bound of the time interval.
- **At risk**: number of individuals that were at risk.

- **Events i**: number of events of type i recorded.
- **All types of events**: number of events of all types recorded.
- **Censored**: number of censored data recorded.
- **Cumulative incidence**: Cumulative incidence obtained for event i at the considered time.
- **Cumulative incidence standard error**: standard error of the previous
- **Cumulative incidence confidence interval**: confidence interval of the previous

Cumulative Survival function: This table displays the various results obtained from the analysis, including:

- **Interval start lime**: lower bound of the time interval.
- **At risk**: number of individuals that were at risk.
- **Events i**: number of events of type i recorded.
- **All types of events**: number of events of all types recorded.
- **Censored**: number of censored data recorded.
- **Cumulative survival function**: Cumulative survival function obtained for event i at the considered time.
- **Cumulative survival function standard error**: standard error of the previous
- **Cumulative survival function confidence interval**: confidence interval of the previous

Charts: Depending on the selected options, up to three charts are displayed: Cumulative incidence and cumulative survival function.

Gray test: For each failure type the Gray test statistic and the associated degrees of freedom and p-values are displayed.

Example

An example of survival analysis based on the cumulative incidence method is available on the XLSTAT Help Center:

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

References

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.

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.

Cox Proportional Hazards Model

Use Cox proportional hazards, also known as Cox regression, to model a survival time using quantitative and/or qualitative covariates.

In this section:

Description

Cox proportional hazards model is a frequently used method in the medical domain (when a patient will get well or not).

The principle of the proportional hazards model is to link the survival time of an individual to covariates. For example, in the medical domain, we are seeking to find out which covariate has the most important impact on the survival time of a patient.

Models

The **Cox model** is a well-recognized statistical technique for exploring the relationship between the survival of a patient and several explanatory variables. The Cox model provides an estimate of the treatment effect on survival after adjustment for other explanatory variables. It allows us to estimate the hazard (or risk) of death, or other event of interest, for individuals, given their prognostic variables.

Interpreting a Cox model involves examining the coefficients for each explanatory variable. A positive regression coefficient for an explanatory variable means that the hazard is higher. Conversely, a negative regression coefficient implies a better prognosis for patients with higher values of that variable.

Cox's method does not assume any particular distribution for the survival times, but it rather assumes that the effects of the different variables on survival are constant over time and are additive in a particular scale.

The hazard function is the probability that an individual will experience an event (for example, death) within a small-time interval, given that the individual has survived up to the beginning of the interval. It can therefore be interpreted as the risk of dying at time t. The hazard function (denoted by $\lambda(t, X)$ can be estimated using the following equation:

$$\lambda(t, X) = \lambda_0(t) \exp(\beta X)$$

The first term only depends on time and the second one depends on X. We are only interested on the second term.

If we only estimate the second term, a very important hypothesis has to be verified: the proportional hazards hypothesis. It means that the hazard ratio between two different observations does not depend on time.

Cox developed a modification of the likelihood function called partial likelihood to estimate the coefficients β not taking into account the time dependant term of the hazard function:

$$\log\left(L(\beta)\right) = \sum_{i=1}^{n} \beta X_i - \log\left(\sum_{j:t_{(j)} \geq t_{(i)}} \exp\left(\beta X_j\right)\right)$$

To estimate the β parameters of the model (the coefficients of the linear function), we try to maximize the partial likelihood function. Contrary to linear regression, an exact analytical solution does not exist. So, an iterative algorithm has to be used. XLSTAT uses a Newton Raphson algorithm. The user can change the maximum number of iterations and the convergence threshold if desired.

Strata

When the proportional hazards hypothesis does not hold, the model can be stratified. If the hypothesis holds on sub-samples, then the partial likelihood is estimated on each sub-sample and these partial likelihoods are summed in order to obtain the estimated partial likelihood. In XLSTAT, strata are defined using a qualitative variable.

Qualitative variables

Qualitative covariates are treated using a complete disjunctive table. In order to have independent variables in the model, the binary variable associated to the first modality of each qualitative variable has to be removed from the model. In XLSTAT, the first modality is always selected and, thus, its effect corresponds to a standard. The impact of the other modalities are obtained relatively to the omitted modality.

Ties handling

The proportional hazards model has been developed by Cox (1972) in order to treat continuous time survival data. However, frequently in practical applications, some observations occur at the

same time. The classical partial likelihood cannot be applied. With XLSTAT, you can use two alternative approaches in order to handle ties:

Breslow's method (1974) (default method): The partial likelihood has the following form:

$$\log\left(L(\beta)\right) = \sum_{i=1}^{T} \beta \sum_{l=1}^{d_i} X_l - d_i \log\left(\sum_{j:t_{(j)} \geq t_{(i)}} \exp\left(\beta X_j\right)\right)$$

where T is the number of times and di is the number of observations associated to time t(i).

Efron's method (1977): The partial likelihood has the following form:

$$\log\left(L(\beta)\right) = \sum_{i=1}^{T} \beta \sum_{l=1}^{d_i} X_l - \sum_{r=0}^{d_i-1} \log\left(\sum_{j:t_{(j)} \geq t_{(i)}} \exp\left(\beta X_j\right) - \frac{r}{d_i} \sum_{j=1}^{d_i} \exp\left(\beta X_j\right)\right)$$

where T is the number of times and d_i is the number of observations associated to time $t_{(i)}$.

If there are no ties, partial likelihoods are equivalent to Cox partial likelihood.

Residuals

Diagnostic procedures for model verification are an important part of a modeling process, and many of these procedures are based on residuals. In survival analysis, and particularly when constructing a Cox proportional hazard model, several types of residuals are used for different purposes.

Martingale residuals are used to examine the overall quality of the fit of a Cox model. They are defined as follows:

$$M_i = d_i - \Lambda_0(t_i)\exp(x_i\beta)$$

Where

$$\Lambda_0 = \sum_{t_i<t} \frac{d_i}{\sum_{j\in R_{t(i)}} \exp(x_j\beta)}$$

is the cumulative risk function.

According to Therneau et al., a problem of these residuals is that they are biased. In addition, they have a maximum value of 1 and a minimum value of $-\infty$

To solve these problems, Therneau et al. have proposed residuals of deviance. They are used for the detection of poorly predicted observations. They are more symmetric around 0 than the Martingale residuals and are defined as follows:

$$D_i = \text{sign}(M_i)\sqrt{-2(M_i + d_i\log(d_i - M_i))}$$

Where M_i is the Martingale residual, and sign represents the sign function.

Observations which correspond to an important residual of deviance are those which are not well adapted to the model, and which can be considered as aberrant observations.

Schoenfeld residuals have been proposed by Schoenfeld as a partial residual which is essential for verifying the proportional hazards hypothesis. He defined these residuals as the difference between the observed value x_{ik} and its conditional expectation knowing the number of individuals still at risk at time t_i : R_i.

$$s_{ik} = d_i(x_{ik} - \bar{x_k})$$

Where

$$\bar{x_k} = \frac{\sum_{j \in R(t_i)} x_{jk} \exp(x\beta)}{\sum_{j \in R(t_i)} \exp(x\beta)}$$

A graph of s_{ik} against t will be centered on 0 if the proportional risks are verified $E(s_{ik}) = 0$.

Score residuals are used to determine influential observations. They are defined as follows:

$$score_{ik} = d_i(x_{ik} - \bar{x_k}(t_i)) - \sum_{t_n < t_i} (x_{ik} - \bar{x_k}(t_n)) \exp(x_i\beta)(\Lambda_0(t_{n-1}))$$

Where

$$\bar{x_k} = \frac{\sum_{j \in R(t_i)} x_{jk} \exp(x\beta)}{\sum_{j \in R(t_i)} \exp(x\beta)}$$

Proportionality test

A central hypothesis of the Cox proportional risk model is that the risk ratio is constant over time. The violation of this hypothesis may lead to a bias in the estimation of the coefficients of the regression. There are different methods to verify if this hypothesis holds.

In XLSTAT, it is possible to verify this hypothesis using the Schoenfeld residuals. Grambsch and Therneau have shown that a normalized version of these residuals approximates the variation of the regression coefficient at time k:$E(s^*_{ik}) + \hat{\beta}_k \approx \beta_k(t_i)$.

Where s^*_{ik} is the normalized Schoenfeld residue of the covariable k at time i : $s^*_{ik} = V^{-1}(\beta, t)S_{ik}$

Where $V(\beta, t)$ is the variance of the vector of estimates β at time t.

Thus, we can test the proportionality of the predictors by creating an interaction with time or a transformation of time (in XLSTAT, the Kaplan-Meier estimator is used as a time transformation). For each variable, we test the nullity of the expectation of the Schoenfeld residuals: H_0 : $\beta_j(t) = \beta_j$ against H_1 : $\beta_j(t) \neq \beta_j$.

A p-value < α indicates a violation of the proportionality hypothesis.

The global test makes it possible to verify this hypothesis for the whole model: $H_0 : \beta(t) = \beta$ against $H_1 : \beta(t) \neq \beta$.

Indices to validate the model

XLSTAT-Life allows you to display indices that help validating the model. They are obtained through bootstraping. As a consequence, for each index you obtain the mean, the standard error, as well as a confidence interval.

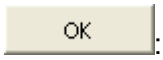
The available indices are:

- R^2(Cox and Snell) : This coefficient, as the classical R^2, takes values between 0 and 1, and measure the goodness of fit of the model. It equals 1 minus the likelihood ratio that compares the likelihood of the model of interest and the likelihood of the independent model;
- R^2(Nagelkerke) : This coefficient, as the classical R^2, takes values between 0 and 1, and measure the goodness of fit of the model. It is equal to the ratio of the Cox and Snell R^2, divided by 1 minus the likelihood of the independent model;
- Shrinkage index: This index allows quantifying the overfitting of the model. When it is lower than 0.85, on can say that there is some overfitting in the model, and that one should reduce the number of parameters in the model.
- The c index: The concordance index (or general discrimination index) allows evaluating the predictive quality of the model. When it is close to 1, the quality is good, and when it is close to 0, it is bad.
- Sommer's D: This index is directly related to the c index, as we have $D = 2 * (c - 0,5)$. As a correlation, it takes values between -1 and 1.

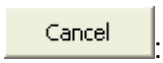
These indices make it easier for the user to validate the Cox model that has been obtained. For a detailed description on the bootstrap and validation for the Cox model, please refer to Harrell *et al.* (1996).

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Date data: Select the data that correspond to the times or the dates when the events or the censoring are recorded. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Status indicator: Select the data that correspond to an event or censoring data. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Event code: Enter the code used to identify an event data within the Status variable. Default value is 1.

Censored code: Enter the code used to identify a censored data within the Status variable. Default value is 0.

Explanatory variables:

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If the variable header has been selected, check that the "Column labels" option has been activated.

Qualitative: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Column labels" option has been activated (see *description*).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selections (time, status and explanatory variables labels) includes a header.

Options tab:

Significance level (%): Enter the significance level for the comparison tests (default value 5%). This value is also used to determine the confidence intervals around the estimated statistics.

Ties handling: Select the method to be used when there is more than one observation for one time (see the description section). Default method: Breslow's method.

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the Newton-Raphson algorithm. The calculations are stopped when the maximum number if iterations has been exceeded. Default value: 100.
- **Convergence**: Enter the maximum value of the evolution of the log of the likelihood from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.

Model selection: Activate this option if you want to use one of the two selection methods provided:

- **Forward**: The selection process starts by adding the variable with the largest contribution to the model. If a second variable is such that its entry probability is greater than the **entry threshold value**, then it is added to the model. This process is iterated until no new variable can be entered in the model.
- **Backward**: This method is similar to the previous one but starts from a complete model.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Test of the null hypothesis $H_0 : \beta = 0$: Activate this option to display the table of statistics associated to the test of the null hypothesis H_0 (likelihood ratio, Wald statistic and score statistic)

Model coefficients: Activate this option to display the table of coefficients for the model. The last columns display the hazard ratios and their confidence intervals (the hazard ratio is calculated as the exponential of the estimated coefficient).

Proportionality test: Activate this option to display the results of the proportionality test.

Predictions: Activate this option to display the prediction vector. These linear predictors are computed as follows: $(x_i - mean(x_i))\beta_i$.

Residuals: Activate this option to display the residuals for all the observations (deviance residuals, martingale residuals, Schoenfeld residuals and score residuals).

Resampled statistics: Activate this option in order to display the validation indexes that have been obtained using the bootstrap method (see the description section).

- **Resamplings**: If the previous option has been activated, enter the number of samples to generate when boostraping.

Charts tab:

Survival distribution function: Activate this option to display the charts corresponding to the cumulative survival distribution function.

-Log(SDF): Activate this option to display the –Log() of the survival distribution function (SDF).

Log(-Log(SDF)): Activate this option to display the Log(–Log()) of the survival distribution function.

Hazard function: Activate this option to display the hazard function when all covariates are at their mean value.

Residuals: Activate this option to display all the residual charts.

Results

XLSTAT displays a large number of tables and charts to help in analyzing and interpreting the results.

Summary statistics: This table displays descriptive statistics for all the variables selected. For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative variables, the categories with their respective frequencies and percentages are displayed.

Summary of the variables selection: When a selection method has been chosen, XLSTAT displays the selection summary.

Goodness of fit coefficients: This table displays a series of statistics for the independent model (corresponding to the case where there is no impact of covariates, beta=0) and for the adjusted model.

- **Observations**: The total number of observations taken into;
- **DF**: Degrees of freedom;
- **-2 Log(Like.)**: The logarithm of the likelihood function associated with the model;
- **AIC**: Akaike's Information Criterion;
- **SBC**: Schwarz's Bayesian Criterion;
- **Iterations**: Number of iterations until convergence.

Test of the null hypothesis $H_0 : \beta = 0$: The H_0 hypothesis corresponds to the independent model (no impact of the covariates). We seek to check if the adjusted model is significantly more powerful than this model. Three tests are available: the likelihood ratio test (-2 Log(Like.)), the Score test and the Wald test. The three statistics follow a χ^2 distribution whose degrees of freedom are shown.

Model parameters: The parameter estimate, corresponding standard deviation, Wald's χ^2, the corresponding p-value and the confidence interval are displayed for each variable of the model. The hazard ratios for each variable with confidence intervals are also displayed.

Proportionality test: For each variable, the correlation between the Schoenfeld residuals and the time vector (1 - Kaplan Meier), the test statistic and its p-value are displayed.

Predictions are given for each observation.

The **residual** table shows, for each observation, the time variable, the censoring variable and the value of the residuals (deviance, martingale, Schoenfeld and score).

Charts: Depending on the selected options, charts are displayed: Cumulative Survival distribution function (SDF), -Log(SDF) and Log(-Log(SDF)), hazard function at mean of covariates, residuals.

Example

A tutorial on how to use Cox regression is available on the XLSTAT Help Center:

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

References

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.

Effron 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.

Proportional Hazards Model with interval censored model

Use the interval-censored proportional hazard model to model survival time based on quantitative or qualitative explanatory variables. This model fits into the framework of survival data methods.

In this section:

Description

The proportional hazard model is one of the most popular regression models for modeling survival times with censored data.

The principle of the proportional hazards model is to link the survival time of an individual to covariates. For example, in the medical domain, we are seeking to find out which covariate has the most important impact on the survival time of a patient.

The first proportional hazard model, introduced by Cox in 1972, works with uncensored data and right censored data. The purpose of the proportional hazard model with interval censored data is therefore the same as for the Cox model, but it will also be possible to model survival times for interval- censored data, uncensored data, left censored data or right censored data.

If the data contains only uncensored or right censored observations, it is possible, with this function, to reproduce the results of a Cox model. However, it is recommended to use Cox's proportional hazards model as it provides a more suitable method for this type of case.

Models

Likelihood

In a proportional hazards model, the survival time of each individual in a population is assumed to follow their own risk function $\lambda_i(t)$ defined by:

$$\lambda_i(t) = \lambda_0(t)\exp(X_i'\beta)$$

Where $\lambda_0(t)$ is the basic hazard function X_i is the vector of the explanatory variables for the i-th individual, and β is the vector of the regression coefficients.

In this model, one will assume that the observations to be analyzed are intervals censored:

$$\{(L_i, R_i], X_i\}_{i=1}^{n}$$

Where L_i and $R_i$represent the left and right limits of the censored interval of the individual i with $L_i \leq R_i$. It can be noted that if $L_i = R_i$ the time of the individual i is uncensored, if $L_i = 0$ the time of the individual i is left censored, and if $R_i = \infty$ the time of the individual i is right censored.

Then, one notes $F(t|X_i)$ the distribution function of the individual i at time t. Under the proportional hazards model, it is given by:

$$F(t|X_i) = 1 - \exp(-\Lambda_0(t)\exp(X_i'\beta))$$

Where $\Lambda_0(t)$ is the cumulative function of the basic hazard and β the regression coefficients.

It is assumed that the time between the beginning of the study and the event is independent of the observation process. Under this assumption, the likelihood function is given by:

$$L = \prod_{i=1}^{n}\{F(R_i|X_i) - F(L_i|X_i)\}$$

If one now distinguishes the three types of censorship, one obtains:

$$L = \prod_{i=1}^{n}\{F(R_i|X_i)^{\delta_1}\{F(R_i|X_i) - F(L_i|X_i)\}^{\delta_2}\{1 - F(L_i|X_i)\}^{\delta_3}$$

Where δ_1 (resp δ_2, δ_3) = 1 if there is left censored (resp interval, right) and 0 otherwise.

Cubic Spline

In the likelihood function given above, the unknown parameters are the regression coefficients β as well as the cumulative function of the basic hazard $\Lambda_0(t)$. One will therefore try to estimate $\Lambda_0(t)$ using I-splines (Ramsay, 1988). This approach gives us the following representation:

$$\Lambda_0(t) = \sum_{l=1}^{k}\gamma_l b_l(t)$$

Where $b_l(t)$ are the basis functions of I-splines, and γ_l are non-zero coefficients which guarantee that $\Lambda_0(t)$ is nondecreasing.

There are different parameters to specify how to build these basis functions of I-splines: the degree of the splines as well as the number and the placement of the knots. As part of this function, the use of cubic splines (order = 3) was favored. As for the number of knots, it is fixed at 3, regularly spaced. It is also possible to choose to optimize this number of knots. In this

case, one will build a model for different numbers of knots and calculate the AIC, the final model will be the one with the lowest AIC. Where AIC is the Akaike's Information Criterion.

EM Algorithm

The EM (Expectation-Maximization) algorithm is an iterative algorithm. It is a parametric estimation method within the framework of maximum likelihood.

It consists of two stages. The first is to calculate the expectation of the log likelihood of the model based on the observed data and the estimation of the parameters $\theta^{(d)} = (\beta^{(d)}, \gamma^{(d)})$. Which give:

$$Q(\theta, \theta^{(d)}) = E[\log(L(\theta))]$$

The second step is to optimize this function. After calculations, one can see that $\theta^{(d+1)}$ is a solution to the system of equations given by:

$$\frac{\partial Q(\theta, \theta^{(d)})}{\partial \beta} = 0$$

And

$$\frac{\partial Q(\theta, \theta^{(d)})}{\partial \gamma_l} = 0$$

Where $l = 1, \ldots, k$.

By solving the 2nd equation, one obtains directly an expression for $\gamma^{(d+1)}$ in terms of $\beta^{(d+1)}$ and the data observed for each l. Thus, by replacing γ_l in $\frac{\partial Q(\theta,\theta^{(d)})}{\partial \beta} = 0$ by the expression of $\gamma^{(d+1)}$, one can obtain $\beta^{(d+1)}$, which allows then the direct computation of $\gamma_l^{(d+1)} = \gamma_l^{(d)}(\beta^{(d+1)})$.

In this function, the resolution of the equation system $\frac{\partial Q(\theta,\theta^{(d)})}{\partial \beta} = 0$ is performed using the Nelder-Mead algorithm, which is an optimization algorithm.

Variance estimation

To compute the information matrix of Fisher $(I(\theta))$, the method of Louis (Louis, 1982) is used. The variance-covariance matrix of $\hat{\theta}$ will then be given by inverting this Fisher information matrix.

Fisher's information matrix is given by:

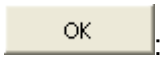
$$I(\theta) = -\frac{\partial^2 Q(\theta, \hat{\theta})}{\partial \theta \partial \hat{\theta}} - var(\frac{\partial \log(L(\theta))}{\partial \theta})$$

Qualitative variables

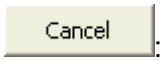
Qualitative covariates are treated using a complete disjunctive table. In order to have independent variables in the model, the binary variable associated to the first modality of each qualitative variable has to be removed from the model. In XLSTAT, we can choose to delete the first or the last modality of each qualitative variable, so the effect of the first or the last modality corresponds to a standard. The impact of the other modalities is related to this omitted modality.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Left endpoint: select here the data corresponding to the left endpoint of the interval. If the data is left censored, put a value of 0. If the data is interval censored, this value must be lower than that of the right endpoint.

Right endpoint: select here the data corresponding to the right endpoint of the interval. If the data is right censored, put a value of 0. If the data is interval censored, this value must be greater than that of the left endpoint.

Explanatory variables:

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The

data selected may be of the numerical type. If the variable header has been selected, check that the "Column labels" option has been activated.

Qualitative: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Column labels" option has been activated (see description).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selections (time, status and explanatory variables labels) includes a header.

Censored code: Select here the vector of the codes of censorship corresponding to the values of the codes entered after.

Uncensored Data Code: Enter the code used to identify uncensored data. The default value is 0.

Left censored code: Enter the code used to identify left censored data. The default value is 1.

Interval Censored Code: Enter the code used to identify interval censored data. The default value is 2.

Right censored code: Enter the code used to identify right censored data. The default value is 3.

Options tab:

Significance level (%): Enter the significance level for the comparison tests (default value 5%). This value is also used to determine the confidence intervals around the estimated statistics.

Optimize number of knots: Enable this option to optimize the number of knots used for spline calculation. The "best" number of knots will then be the one that optimizes the AIC of the model. In case this option is not activated, the number of nodes will be 3.

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the Newton-Raphson algorithm. The calculations are stopped when the maximum number if iterations has been exceeded. Default value: 100.

- **Convergence**: Enter the maximum value of the evolution of the log of the likelihood from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.

- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Model coefficients: Activate this option to display the table of coefficients for the model. The last columns display the hazard ratios and their confidence intervals (the hazard ratio is calculated as the exponential of the estimated coefficient).

Predictions: Activate this option to display the prediction vector. These linear predictors are computed as follows: $(x_i - mean(x_i))\beta_i$

Charts tab:

Survival distribution function: Activate this option to display the charts corresponding to the cumulative survival distribution function.

-Log(SDF): Activate this option to display the –Log() of the survival distribution function (SDF).

Log(-Log(SDF)): Activate this option to display the Log(–Log()) of the survival distribution function.

Results

XLSTAT displays a large number of tables and charts to help in analyzing and interpreting the results.

Summary statistics: This table displays descriptive statistics for all the variables selected. For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative variables, the categories with their respective frequencies and percentages are displayed.

Goodness of fit coefficients: This table displays a series of statistics for the independent model (corresponding to the case where there is no impact of covariates, beta=0) and for the adjusted model.

- **Observations**: The total number of observations taken into;
- **DF**: Degrees of freedom;
- **-2 Log(Like.)**: The logarithm of the likelihood function associated with the model;
- **AIC**: Akaike's Information Criterion;
- **SBC**: Schwarz's Bayesian Criterion;
- **Iterations**: Number of iterations until convergence.

Model parameters: The parameter estimate, corresponding standard deviation, Wald's χ^2, the corresponding p-value and the confidence interval are displayed for each variable of the model. The hazard ratios for each variable with confidence intervals are also displayed.

Predictions are given for each observation.

Charts: Depending on the selected options, charts are displayed: Cumulative Survival distribution function (SDF), -Log(SDF) and Log(-Log(SDF)).

Example

A tutorial on how to use proportional hazards model with interval censored data is available on the XLSTAT Help Center:

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

References

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). Regression 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.

Parametric survival models

Use the parametric survival model, also known as Weibull model, to model a survival time using a given probability distribution and, if necessary, quantitative and/or qualitative covariates. These models fit into the framework of the methods for survival data analysis.

In this section:

Description

The parametric survival model is a method that applies in the context of the analysis of survival data. It allows modelling survival time with right- censored data. It is widely used in medicine (survival time or cure of a patient).

The principle of the parametric survival model is to link the survival time of an individual to a probability distribution (the Weibull distribution is often used) and, when necessary, covariates. For example, in the medical domain, we are seeking to find out which covariate has the most important impact on the survival time of a patient based on a defined distribution.

XLSTAT-Life offers two tools for parametric survival models:

- The **parametric survival regression**, which lets you apply a regression model and analyze the impact of explanatory variables on survival time (assuming an underlying distribution).

- The **parametric survival curve** uses a chosen distribution to model the survival time.

These two methods are exactly equivalent to a methodological standpoint, the difference lies in the fact that, in the first case, explanatory variables are included.

Models

The parametric survival model is similar to the classical regression models in the sense that one tries to link an event (modelled by a date) to a number of explanatory variables.

The parametric survival model is a parametric model. It is based on the assumption that survival times follow a distribution. This assumes a structure for the hazard function that is associated with the chosen distribution.

The parametric survival model is applicable to any situation where one which to study the time of occurrence of an event. This event may be the recurrence of a disease, the response to a

treatment, the death, etc. For each subject, we know the date of the latest event (censored or not).

The subjects for which we do not know the status are censored data. The explanatory variables are noted X_j and do not vary along the study.

The T variable is the time until the event. The parametric survival model can express the risk of occurrence of the event as a function of time t and of the explanatory variables X_j. These variables may represent risk factors, prognostic factors, treatment, about the intrinsic characteristics, ...

The survival function, noted $S(t)$, is defined depending on the selected distribution. XLSTAT-Life offers different distributions, among others, the exponential distribution (the survival rate is constant, $h(t) = l$), the Weibull distribution (often called Weibull model), the distributions of extreme values?...

The exponential and Weibull models are very interesting because they are simultaneously proportional hazards models (such as the Cox model) and accelerated failure time models (for all individuals i and j with survival time $S_i()$ and $S_j()$, there exists a constant ϕ such that $S_i(t) = S_j(t * \phi)$ for all t).

The estimation of such models is done with the maximum likelihood method. Generally $Y = log(T)$ is used as dependent variable (for Weibull and exponential models).

Unlike linear regression, an exact analytical solution does not exist. It is therefore necessary to use an iterative algorithm. XLSTAT uses a Newton- Raphson algorithm. The user can change if desired maximum number of iterations and the convergence threshold.

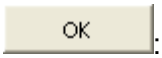
Interpretation of results is done both by studying the graphs associated with cumulative survival functions and studying the tables of coefficients and goodness of fit indices.

Qualitative variables

Qualitative covariates are treated using a complete disjunctive table. In order to have independent variables in the model, the binary variable associated to the first modality of each qualitative variable has to be removed from the model. In XLSTAT, the first or the last modality can be selected and, thus, its effect corresponds to a standard. The impacts of the other modalities are obtained relatively to the omitted modality.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Date data: Select the data that correspond to the times or the dates when the events or the censoring are recorded. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Status indicator: Select the data that correspond to an event or censoring data. If a column header has been selected on the first row, check that the "Column labels" option has been activated.

Event code: Enter the code used to identify an event data within the Status variable. Default value is 1.

Censored code: Enter the code used to identify a censored data within the Status variable. Default value is 0.

Explanatory variables (in the case of a parametric survival regression):

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If the variable header has been selected, check that the "Column labels" option has been activated.

Qualitative: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Column labels" option has been activated (see *description*).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selections (time, status and explanatory variables labels) includes a header.

Distribution: Select the distribution to be used to fit your model. XLSTAT-Life offers different distributions including Weibull, exponential, extreme value…

Regression weights: Activate this option if you want to carry out a weighted least squares regression. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Significance level (%): Enter the significance level for the comparison tests (default value 5%). This value is also used to determine the confidence intervals around the estimated statistics.

Initial parameters: Activate this option if you want to take initial parameters into account. If you do not activate this option, the initial parameters are automatically obtained. If a column header has been selected, check that the "Variable labels" option is activated.

Fixed constant: Activate this option to fix the constant of the regression model to a value you then enter (0 by default).

Tolerance: Activate this option to prevent the initial regression calculation algorithm taking into account variables which might be either constant or too correlated with other variables already used in the model (0.0001 by default).

Constraints: Details on the various options are available in the description section.

a1 = 0: Choose this option so that the parameter of the first category of each factor is set to 0.

an = 0: Choose this option so that the parameter of the last category of each factor is set to 0.

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the Newton-Raphson algorithm. The calculations are stopped when the maximum number if iterations has been exceeded. Default value: 100.
- **Convergence**: Enter the maximum value of the evolution of the log of the likelihood from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.

Model selection: Activate this option if you want to use one of the two selection methods provided:

- **Forward**: The selection process starts by adding the variable with the largest contribution to the model. If a second variable is such that its entry probability is greater than the **entry**

threshold value, then it is added to the model. This process is iterated until no new variable can be entered in the model.

- **Backward**: This method is similar to the previous one but starts from a complete model.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Test of the null hypothesis H0: beta=0: Activate this option to display the table of statistics associated to the test of the null hypothesis H0 (likelihood ratio, Wald statistic and score statistic)

Model coefficients: Activate this option to display the table of coefficients for the model. The last columns display the hazard ratios and their confidence intervals (the hazard ratio is calculated as the exponential of the estimated coefficient).

Residuals and predictions: Activate this option to display the residuals for all the observations (standardized residuals, Cox-Snell residuals). The value of the estimated cumulative distribution function, the hazard function and the cumulative survival function for each observation are displayed.

Quantiles: Activate this option to display the quantiles for each observation (in the case of a parametric survival regression) and for different values of the percentiles (1, 5, 10, 25, 50, 75, 90, 95 and 99 %)..

Charts tab:

Survival distribution function: Activate this option to display the charts corresponding to the cumulative survival distribution function.

-Log(SDF): Activate this option to display the –Log() of the survival distribution function (SDF).

Log(-Log(SDF)): Activate this option to display the Log(–Log()) of the survival distribution function.

Hazard function: Activate this option to display the hazard function when all covariates are at their mean value.

Residuals: Activate this option to display all the residual charts.

Results

XLSTAT displays a large number of tables and charts to help in analyzing and interpreting the results.

Summary statistics: This table displays descriptive statistics for all the variables selected. For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative variables, the categories with their respective frequencies and percentages are displayed.

Summary of the variables selection: When a selection method has been chosen, XLSTAT displays the selection summary.

Goodness of fit coefficients: This table displays a series of statistics for the independent model (corresponding to the case where there is no impact of covariates, beta=0) and for the adjusted model.

- **Observations**: The total number of observations taken into;
- **DF**: Degrees of freedom;
- **-2 Log(Like.)**: The logarithm of the likelihood function associated with the model;
- **AIC**: Akaike's Information Criterion;
- **SBC**: Schwarz's Bayesian Criterion;
- **Iterations**: Number of iterations until convergence.

Test of the null hypothesis H0: beta=0: The H0 hypothesis corresponds to the independent model (no impact of the covariates). We seek to check if the adjusted model is significantly more powerful than this model. Three tests are available: the likelihood ratio test (-2 Log(Like.)), the Score test and the Wald test. The three statistics follow a Chi² distribution whose degrees of freedom are shown.

Model parameters: The parameter estimate, corresponding standard deviation, Wald's Chi², the corresponding p-value and the confidence interval are displayed for each variable of the model. Confidence intervals are also displayed.

The **residual and predictions** table shows, for each observation, the time variable, the censoring variable, the value of the residuals, the cumulative distribution function, the cumulative survival function and the hazard function..

Charts: Depending on the selected options, charts are displayed: Cumulative Survival distribution function (SDF), -Log(SDF) and Log(-Log(SDF)), hazard function, residuals.

Example

A tutorial on how to use parametric survival regression is available on the XLSTAT Help Center:

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

A tutorial on how to use parametric survival curve is available on the XLSTAT Help Center:

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

References

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.

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. 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.

Propensity score matching

Use the propensity score matching to match participants of two distinct groups in order to control the effect of confounding variables in observational studies.

In this section:

Description

The propensity score is defined as the probability for a participant to belong to one of two groups given some variables known as confounders. The propensity score matching is a technique that attempts to reduce the possible bias associated to those confounding variables in observational studies.

Presentation

A typical example where confounding variables might be encountered would be a study aiming to evaluate the effect of a new drug. Participants would belong to the treated group if they received the new drug and to the control group if they did not. The study would then measure the survival rate for each group after a certain amount of time. If the survival rate of the treated group is found to be lower to the one of the control group, one might be tempted to conclude that the drug doesn't bring any benefit at all or, worse, that it has some dangerous effect on the health of the participants.

In fact, the two groups are not identical and the new drug was administrated to a group of people that already had a serious disease diagnosed before the study began. On the contrary, the control group was made up of a relatively healthy group of people where only a small proportion had a serious disease detected.

In this example, the variable "serious disease detected" is a confounding variable because its value has an effect on the probability of a given participant to belong to one group or the other. Controlling the effect of this confounding variable is highly recommended as it may otherwise introduce a serious bias in the experimental results. In our case, participants of the treated group had a serious disease detected which could explain a higher mortality rate for this group compared to the control group.

Confounding variables are inherent to studies where the assignment procedure of a participant to one group is not random. Possible reasons to use non- random procedures are numerous,

they might be ethical, legal, economical or simply practical. The study is then defined as an observational study or a non-randomized trial.

Propensity score matching is one of the most successful techniques that aim to minimize the effect of confounding variables. The basic idea is to estimate the probability of a given participant to belong to one group by adjusting a logistic regression model on the group variable with the suspected confounding variables as predictors. This probability, called the propensity score, is expected to reflect the effect of the identified confounders.

The propensity score is then used to match each participant of the treatment group to the most similar participant of the control group. Finally, matched participants form two groups that are more comparable in term of confounding variables and possible biases in the experimental results are expected to be reduced.

Estimating the propensity score

The propensity score was first introduced in Rosenbaum P.R., Rubin D.B. (1983a) as the treatment assignment conditional on covariates:

$$p_i = \Pr(Z_i = 1|X_i)$$

Where p_i is the propensity score, Z_i indicates the group (treatment or control) and X_i refers to the covariates or suspected confounding variables.

In XLSTAT, the propensity score is estimated using a logistic regression or logit model (see also Logistic Regression for a more detailed description).

The group variable is the dependant variable of the logit model. It is a binary variable that separates participants that belong to the treatment group from those who belong to the control group. You may choose which one of the two categories corresponds to the treatment group.

The explanatory variables of the logit model are the confounding variables of our propensity score model. They might be continuous (quantitative) or categorical (qualitative) data and XLSTAT allows you to estimate interactions between those variables.

Matching algorithm

Once the propensity score has been estimated, each participant of the treatment group is matched to the most similar participant of the control group (Rosenbaum P. R. (1989)**)**. The similarity is evaluated as a distance on the logit function of the propensity score defined as:

$$\text{logit}(p_i) = \log(\frac{p_i}{1-p_i})$$

The distance matrix is computed between the treatment group and the control group. XLSTAT implementation proposes two metrics:

- the Euclidean distance
- the Mahalanobis distance

The user can also set an upper limit above which a distance between two participants is considered too high for a match to be possible. We talk of caliper radius to refer to this limit and it is defined in XLSTAT as:

$$C = a\sqrt{\frac{\sigma_T^2 + \sigma_C^2}{2}}$$

Where C is the caliper radius, σ_T^2 the variance of $\text{logit}(p_i)$ of the treatment group, σ_C^2 the variance of $\text{logit}(p_i)$ of the control group and α a coefficient. No strong consensus exists in the literature about the value α should be given. Most frequent values are 0.1, 0.2 and 0.25. Those values and some less typical values (0.5 and 1) are proposed in XLSTAT. The user is also given the possibility to set C to the value of his choice.

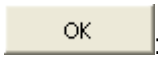
Two algorithms are available in XLSTAT to perform the matching operation: the greedy algorithm and the optimal algorithm. With both of these algorithms, it is possible to match each participant of the treatment group to one participant of the control group, to a specified number of participants of the control group or to all participants of the control group. The latter two configurations can lead to substantially different results depending on the chosen matching algorithm.

The greedy algorithm proceeds by successively matching each participant of the treatment group to the best candidate among the available participants in the control group. By available, it is meant that candidates should be at a distance smaller than the caliper radius (if the caliper option is activated) and that the greedy algorithm performs matches without replacement. Once a participant of the control group has been matched to one participant of the treatment group, it is not available anymore for the subsequent matches. Therefore, the initial order of the participants of the treatment group has an effect on the final result of the matching operation. To mitigate this aspect of the greedy algorithm, XLSTAT proposes a random shuffling of participants that can be activated by the user.

On the contrary, the optimal algorithm does not suffer this type of limitation. It is based on a minimum-cost flow optimization that will minimize the total distance for all participants of the matching operation. The matching operation remains without replacement but the order is not an issue anymore. With this algorithm, an additional feature is available to balance the matching operation according to a given qualitative variable. This feature should be used if the frequency of occurrence of each category of the balancing variable in the treatment group is expected to be reproduced in the matched control group.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.

: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Group variable: Select the group variable you want to model. This should be a binary variable indicating if a participant belongs to the group that received the treatment or to the control group. If a column header has been selected, check that the "Variable labels" option has been activated.

Treatment modality: Select the modality of the group variable that corresponds to the group that received the treatment. Modalities should be detected automatically and proposed in the drop-down menu when the group variable is selected. If detection fails or if the drop-down list does not correspond to the variable after the selection was changed, you can click on the refresh button just at the right to refresh the display.

Explanatory variables:

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be numerical. If the variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Options tab:

Tolerance: Enter the value of the tolerance threshold below which a variable will automatically be ignored.

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95.

Interactions / Level: Activate this option to include interactions in the model then enter the maximum interaction level (value between 1 and 4).

Firth's method: Activate this option to use Firth's penalized likelihood (see description). This option is only available for binary logit model.

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the Newton-Raphson algorithm. The calculations are stopped when the maximum number if iterations has been exceeded. Default value: 100.
- **Convergence**: Enter the maximum value of the evolution of the log of the likelihood from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.

Shuffle rows: Activate this option to shuffle rows (participants) before performing the greedy matching procedure. This option is only available when the greedy algorithm has been selected.

Balance groups: Activate this option if you wish to use a categorical variable to balance some groups between the treatment group and the control group in the matching process. Then select the corresponding variable in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Matching method:

Greedy algorithm: Activate this option if you want to use the greedy algorithm during the matching process.

Optimal algorithm: Activate this option if you want to use the optimal algorithm during the matching process.

Euclidean / Mahalanobis distance: Select the type of distance you want to use perform the matching operation.

Number of matches:

One to one: Activate this option to match each participant of the treatment group to a participant of the control group provided that there are suitable candidates.

One to several: Activate this option to match each participant of the treatment group to a given number of participants of the control group provided that there are suitable candidates. Then enter the desired number of participants.

One-to-all: Activate this option to match every participant of the control group to one of the participant of the treatment group.

Caliper: Activate this option to set a limit on the distance acceptable for a match. Then select the the caliper radius you want using the drop-down menu. If you select "User defined" you also have to enter the desired value.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Type II analysis: Activate this option to display the type II analysis of variance table.

Hosmer-Lemeshow test: Activate this option to display the results of the Hosmer-Lemeshow test.

Model coefficients: Activate this option to display the table of coefficients for the model. Optionally, **confidence intervals** of type "*profile likelihood* " can be calculated (see description).

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Summary of matched observations: Activate this option if you want a summary of the matched observations to be displayed.

Propensity score: Activate this option if you want the list of all computed propensity scores to be displayed.

Distance matrix: Activate this option if you want the distance matrix to be displayed.

Detailed matched observations: Activate this option if you want the matched observations to be displayed.

Charts tab:

Standardized coefficients: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.

ROC curve: Activate this option to display the ROC curve.

Box plot of scores: Activate this option to display the box plot of scores for each group.

Results

XLSTAT displays a large number of tables and charts to help in analyzing and interpreting the results.

Summary statistics: This table displays descriptive statistics for all the variables selected. For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative variables, the categories with their respective frequencies and percentages are displayed.

Goodness of fit coefficients: This table displays a series of statistics for the independent model (corresponding to the case where the linear combination of explanatory variables reduces to a constant) and for the adjusted model.

- **Observations**: The total number of observations taken into account (sum of the weights of the observations);
- **Sum of weights**: The total number of observations taken into account (sum of the weights of the observations multiplied by the weights in the regression);
- **DF**: Degrees of freedom;
- **-2 Log(Like.)**: The logarithm of the likelihood function associated with the model;
- **R^2 (McFadden):** Coefficient, like the R^{2}^, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model;

- **R²(Cox and Snell):** Coefficient, like the R²^, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model raised to the power 2/Sw, where Sw is the sum of weights.
- **R²(Nagelkerke):** Coefficient, like the R²^, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to ratio of the R² of Cox and Snell, divided by 1 minus the likelihood of the independent model raised to the power 2/Sw;
- **AIC**: Akaike's Information Criterion;
- **SBC**: Schwarz's Bayesian Criterion.
- **Iterations**: Number of iterations before convergence.

Test of the null hypothesis H0: Y=p0: The H0 hypothesis corresponds to the independent model which gives probability p0 whatever the values of the explanatory variables. We seek to check if the adjusted model is significantly more powerful than this model. Three tests are available: the likelihood ratio test (-2 Log(Like.)), the Score test and the Wald test. The three statistics follow a Chi²^ distribution whose degrees of freedom are shown.

Type II analysis: This table is only useful if there is more than one explanatory variable. Here, the adjusted model is tested against a test model where the variable in the row of the table in question has been removed. If the probability Pr > LR is less than a significance threshold which has been set (typically 0.05), then the contribution of the variable to the adjustment of the model is significant. Otherwise, it can be removed from the model.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can easily be seen on the chart of standardized coefficients), the weight of a variable in the model is not significant.

ROC curve: The ROC curve is used to evaluate the performance of the model by means of the area under the curve (AUC) and to compare several models together (see the description section for more details).

The table of **Summary matching** display some indicators on the proportion of participants that have been matched. The total cost in term of distance is also given just below the table.

The table of **propensity scores** gives the calculated propensity score for each participant of the two groups. The value of the logit of the propensity score is also given. This is the value that is used to compute the distance between each participant. Upper and lower bounds are also given for the two variables.

The **distance matrix** is also displayed to give a general view of all the computed distances. Participants of the treatment group are on rows, those of the control group are on columns. Distances for match pairs are displayed in bold.

The **Box plot** displays several parameters on the logit distribution of propensity scores for the complete treatment group, the portion of the treatment group that has been matched, the portion of the control group that has been matched and the complete control group.

Example

A tutorial on how to use propensity score matching is available on the XLSTAT Help Center:

https://help.xlstat.com/customer/en/portal/articles/2826861-propensity-score-matching-in-excel?b_id=9283

References

Rosenbaum P.R. and 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*, **84(408)**, 1024-1032

Sensitivity and Specificity

Use this tool to compute, among others, the sensitivity, specificity, odds ratio, predictive values, and likelihood ratios associated with a test or a detection method. These indices can be used to assess the performance of a test. For example in medicine it can be used to evaluate the efficiency of a test used to diagnose a disease or in quality control to detect the presence of a defect in a manufactured product.

In this section:

Description

This method was first developed during World War II to develop effective means of detecting Japanese aircrafts. It was then applied more generally to signal detection and medicine where it is now widely used.

The problem is as follows: we study a phenomenon, often binary (for example, the presence or absence of a disease) and we want to develop a test to detect effectively the occurrence of a precise event (for example, the presence of the disease).

Let V be the binary or multinomial variable that describes the phenomenon for N individuals that are being followed. We note by + the individuals for which the event occurs and by ‘-those for which it does not. Let T be a test which goal is to detect if the event occurred or not. T can be a binary (presence/absence), a qualitative (for example the color), or a quantitative variable (for example a concentration). For binary or qualitative variables, let t1 be the category corresponding to the occurrence of the event of interest. For a quantitative variable, let t1 be the threshold value under or above which the event is assumed to happen.

Once the test has been applied to the N individuals, we obtain an individuals/variables table in which for each individual you find if the event occurred or not, and the result of the 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

Case of binary test Case of a quantitative test

These tables can be summarized in a 2x2 contingency table:

	M+	**M-**
T+	25	12
T-	8	13

In the example above, there are 25 individuals for whom the test has detected the presence of the disease and 13 for which it has detected its absence. However, for 20 individuals diagnosis is bad because for 8 of them the test contends the absence of the disease while the patients are sick, and for 12 of them, it concludes that they are sick while they are not.

The following vocabulary is being used:

True positive (TP): Number of cases that the test declares positive and that are truly positive.

False positive (FP): Number of cases that the test declares positive and that in reality are negative.

True n egati ve (VN): Number of cases that the test declares negative and that are truly negative.

False negative (FN): Number of cases that the test declares negative and that in reality are positive.

Several indices have been developed to evaluate the performance of a test:

Sensitivity (equivalent to the **True Positive Rate**): Proportion of positive cases that are well detected by the test. In other words, the sensitivity measures how the test is effective when used on positive individuals. The test is perfect for positive individuals when sensitivity is 1, equivalent to a random draw when sensitivity is 0.5. If it is below 0.5, the test is counter-performing and it would be useful to reverse the rule so that sensitivity is higher than 0.5 (provided that this does not affect the specificity). The mathematical definition is given by: Sensitivity = TP/(TP + FN).

Specificity (also called **True Negative Rate**): proportion of negative cases that are well detected by the test. In other words, specificity measures how the test is effective when used on negative individuals. The test is perfect for negative individuals when the specificity is 1, equivalent to a random draw when the specificity is 0.5. If it is below 0.5, the test is counter

performing-and it would be useful to reverse the rule so that specificity is higher than 0.5 (provided that this does not affect the sensitivity). The mathematical definition is given by: Specificity = TN/(TN + FP).

False Positive Rate (FPR): Proportion of negative cases that the test detects as positive (FPR = 1-Specificity).

False Negative Rate (FNR): Proportion of positive cases that the test detects as negative (FNR = 1-Sensitivity)

Prevalence: relative frequency of the event of interest in the total sample (TP+FN)/N.

Positive Predictive Value (PPV): Proportion of truly positive cases among the positive cases detected by the test. We have PPV = TP / (TP + FP), or PPV = Sensitivity x Prevalence / [(Sensitivity x Prevalence + (1-Specificity)(1-Prevalence)]. It is a fundamental value that depends on the prevalence, an index that is independent of the quality of the test.

Negative Predictive Value (NPV): Proportion of truly negative cases among the negative cases detected by the test. We have NPV = TN / (TN + FN), or PPV = Specificity x (1 - Prevalence) / [(Specificity (1-Prevalence) + (1-Sensibility) x Prevalence]. This index depends also on the prevalence that is independent of the quality of the test.

Positive Likelihood Ratio (LR+): This ratio indicates to which point an individual has more chances to be positive in reality when the test is telling it is positive. We have LR+ = Sensitivity / (1-Specificity). The LR+ is a positive or null value.

Negative Likelihood Ratio (LR-): This ratio indicates to which point an individual has more chances to be negative in reality when the test is telling it is positive. We have LR- = (1-Sensitivity) / (Specificity). The LR- is a positive or null value.

Odds ratio: The odds ratio indicates how much an individual is more likely to be positive if the test is positive, compared to cases where the test is negative. For example, an odds ratio of 2 means that the chance that the positive event occurs is twice higher if the test is positive than if it is negative. The odds ratio is a positive or null value. We have Odds ratio = TPxTN / (FPxFN).

Relative risk: The relative risk is a ratio that measures how better the test behaves when it is a positive report than when it is negative. For example, a relative risk of 2 means that the test is twice more powerful when it is positive that when it is negative. A value close to 1 corresponds to a case of independence between the rows and columns, and to a test that performs as well when it is positive as when it is negative. Relative risk is a null or positive value given by: Relative risk = TP/(TP+FP) / (FN/(FN+TN)).

Confidence intervals

For the various presented above, several methods of calculating their variance and, therefore their confidence intervals, have been proposed. There are two families: the first concerns proportions, such as sensitivity and specificity, and the second ratios, such as LR +, LR- the odds ratio and the relative risk.

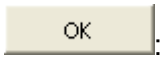
For proportions, XLSTAT allows you to use the simple (Wald, 1939) or adjusted (Agresti and Coull, 1998) Wald intervals, a calculation based on the Wilson score (Wilson, 1927), possibly with a correction of continuity, or the Clopper-Pearson (1934) intervals. Agresti and Caffo recommend using the adjusted Wald interval or the Wilson score intervals.

For ratios, the variances are calculated using a single method, with or without correction of continuity.

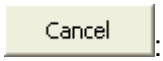
Once the variance of the above statistics is calculated, we assume their asymptotic normality (or of their logarithm for ratios) to determine the corresponding confidence intervals. Many of the statistics are proportions and should lie between 0 and 1. If the intervals fall partly outside these limits, XLSTAT automatically corrects the bounds of the interval.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data format:

2x2 table (Test/Event): Choose this option if your data are available in a 2x2 contingency table with the tests results in rows and the positive and negative events in columns. You can then specify in which column of the table are the positive events, and on which row are the cases detected as positive by the test. The option "Labels included" must be activated if the labels of the rows and columns were selected with the data.

Individual data: Choose this option if your data are recorded in a individuals/variables table. You must then select the **event data** that correspond to the phenomenon of interest (for example, the presence or absence of a disease) and specify which **code** is associated with positive events (for example + when a disease is diagnosed). You must also select the **test data** corresponding to the value of the diagnostic test. This test may be quantitative (concentration), binary (positive or negative) or qualitative (color). If the test is quantitative, you must specify if XLSTAT should consider it as positive when the test is above or below a given threshold value. If the test is qualitative or binary, you must select the value corresponding to a positive test.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if the row and column labels are selected. This option is available if you selected the "2x2 table" format.

Variable labels: Activate this option if, in column mode, the first row of the selected data contains a header, or in row mode, if the first column of the selected data contains a header. This option is available if you selected the "individual data" format.

Weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Confidence intervals:

- **Size (%)**: Enter the size of the confidence interval in % (default value: 95).
- **Wald**: Activate this option if you want to calculate confidence intervals on the various indexes using the approximation of the binomial distribution by the normal distribution. Activate "Adjusted" to use the adjustment of Agresti and Coull.
- **Wilson score**: Activate this option if you want to calculate confidence intervals on the various indexes using the Wilson score approximation.
- **Clopper-Pearson**: Activate this option if you want to calculate confidence intervals on the various indexes using the Clopper-Pearson approximation.
- **Continuity correction**: Activate this option if you want to apply the continuity correction to the Wilson score and to the interval on ratios.

A priori prevalence: If you know that the disease involves a certain proportion of individuals in the total population, you can use this information to adjust predictive values calculated from your sample.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Results

The results are made of the contingency table followed by the table that displays the various indices described in the description section.

Example

An example showing how to compute sensitivity and specificity is available on the XLSTAT Help Center:

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

References

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.

ROC curves

Use this tool to generate an ROC curve that allows to represent the evolution of the proportion of true positive cases (also called sensitivity) as a function of the proportion of false positives cases (corresponding to 1 minus specificity), and to evaluate a binary classifier such as a test to diagnose a disease, or to control the presence of defects on a manufactured product.

In this section:

Description

ROC curves have first been developed during World War II to develop effective means of detecting Japanese aircrafts. This methodology was then applied more generally to signal detection and medicine where it is now widely used.

The problem is as follows: we study a phenomenon, often binary (for example, the presence or absence of a disease) and we want to develop a test to effectively detect the occurrence of a precise event (for example, the presence of the disease).

Let V be the binary or multinomial variable that describes the phenomenon for N individuals that are being followed. We note by + the individuals for which the event occurs and by - those for which it does not. Let T be a test which goal is to detect if the event occurred or not. T is most of the time continuous (for example, a concentration) but it can also be ordinal (to represent levels).

We want to set the threshold value below or beyond which the event occurs. To do so, we examine a set of possible threshold values for each we calculate various statistics among which the simplest are:

- **True positive (TP)**: Number of cases that the test declares positive and that are truly positive.
- **False positive (FP)**: Number of cases that the test declares positive and that in reality are negative.
- **True negative (VN)**: Number of cases that the test declares negative and that are truly negative.
- **False negative (FN)**: Number of cases that the test declares negative and that in reality are positive.

- **Prevalence**: Relative frequency of the event of interest in the total sample (TP+FN)/N.

Several indices have been developed to evaluate the performance of a test at a given threshold value:

Sensitivity (equivalent to the **True Positive Rate**): Proportion of positive cases that are well detected by the test. In other words, the sensitivity measures how the test is effective when used on positive individuals. The test is perfect for positive individuals when sensitivity is 1, equivalent to a random draw when sensitivity is 0.5. If it is below 0.5, the test is counter-performing and it would be useful to reverse the rule so that sensitivity is higher than 0.5 (provided that this does not affect the specificity). The mathematical definition is given by: Sensitivity = TP/(TP + FN).

Specificity (also called **True Negative Rate**): Proportion of negative cases that are well detected by the test. In other words, specificity measures how the test is effective when used on negative individuals. The test is perfect for negative individuals when the specificity is 1, equivalent to a random draw when the specificity is 0.5. If it is below 0.5, the test is counter-performing and it would be useful to reverse the rule so that specificity is higher than 0.5 (provided that this does not affect the sensitivity). The mathematical definition is given by: Specificity = TN/(TN + FP).

False Positive Rate (FPR): Proportion of negative cases that the test detects as positive (FPR = 1-Specificity).

False Negative Rate (FNR): Proportion of positive cases that the test detects as negative (FNR = 1-Sensitivity)

Prevalence: Relative frequency of the event of interest in the total sample (TP+FN)/N.

Positive Predictive Value (PPV): Proportion of truly positive cases among the positive cases detected by the test. We have PPV = TP / (TP + FP), or PPV = Sensitivity x Prevalence / [(Sensitivity x Prevalence + (1-Specificity)(1-Prevalence)]. It is a fundamental value that depends on the prevalence, an index that is independent of the quality of the test.

Negative Predictive Value (NPV): Proportion of truly negative cases among the negative cases detected by the test. We have NPV = TN / (TN + FN), or PPV = Specificity x (1 - Prevalence) / [(Specificity (1-Prevalence) + (1-Sensibility) x Prevalence]. This index depends also on the prevalence that is independent of the quality of the test.

Positive Likelihood Ratio (LR+): This ratio indicates to which point an individual has more chances to be positive in reality when the test is telling it is positive. We have LR+ = Sensitivity / (1-Specificity). The LR+ is a positive or null value.

Negative Likelihood Ratio (LR-): This ratio indicates to which point an individual has more chances to be negative in reality when the test is telling it is positive. We have LR- = (1-Sensitivity) / (Specificity). The LR- is a positive or null value.

Odds ratio: The odds ratio indicates how much an individual is more likely to be positive if the test is positive, compared to cases where the test is negative. For example, an odds ratio of 2 means that the chance that the positive event occurs is twice higher if the test is positive than if it is negative. The odds ratio is a positive or null value. We have Odds ratio = TPxTN / (FPxFN).

Relative risk: The relative risk is a ratio that measures how better the test behaves when it is a positive report than when it is negative. For example, a relative risk of 2 means that the test is twice more powerful when it is positive that when it is negative. A value close to 1 corresponds to a case of independence between the rows and columns, and to a test that performs as well when it is positive as when it is negative. Relative risk is a null or positive value given by: Relative risk = TP/(TP+FP) / (FN/(FN+TN)).

Confidence intervals

For the various presented above, several methods of calculating their variance and, therefore their confidence intervals, have been proposed. There are two families: the first concerns proportions, such as sensitivity and specificity, and the second ratios, such as LR +, LR- the odds ratio and the relative risk.

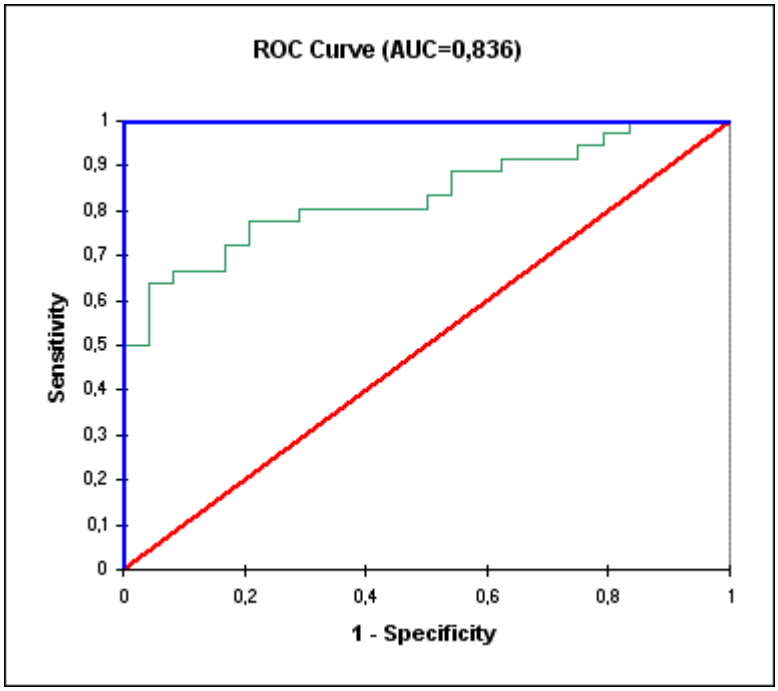
For proportions, XLSTAT allows you to use the simple (Wald, 1939) or adjusted (Agresti and Coull, 1998) Wald intervals, a calculation based on the Wilson score (Wilson, 1927), possibly with a correction of continuity, or the Clopper-Pearson (1934) intervals. Agresti and Caffo recommend using the adjusted Wald interval or the Wilson score intervals.

For ratios, the variances are calculated using a single method, with or without correction of continuity.

Once the variance of the above statistics is calculated, we assume their asymptotic normality (or of their logarithm for ratios) to determine the corresponding confidence intervals. Many of the statistics are proportions and should lie between 0 and 1. If the intervals fall partly outside these limits, XLSTAT automatically corrects the bounds of the interval.

ROC curve

The ROC curve corresponds to the graphical representation of the couple (1 – specificity, sensitivity) for the various possible threshold values.



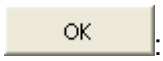
The area under the curve (AUC) is a synthetic index calculated for ROC curves. The AUC is the probability that a positive event is classified as positive by the test given all possible values of the test. For an ideal model we have AUC = 1 (above in blue), where for a random pattern we have AUC = 0.5 (above in red). One usually considers that the model is good when the value of the AUC is higher than 0.7. A well discriminating model should have an AUC between 0.87 and 0.9. A model with an AUC above 0.9 is excellent.

Sen (1960), Bamber (1975) and Hanley and McNeil (1982) have proposed different methods to calculate the variance of the AUC. All are available in XLSTAT. XLSTAT offers a comparison test of the AUC as well to 0.5, the value 0.5 corresponding to a random classifier. This test is based on the difference between the AUC and 0.5 divided by the variance calculated according to one of the three proposed methods. The statistic obtained is supposed to follow a standard normal distribution, which allows the calculation of the p-value.

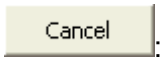
The AUC can also be used to compare different tests between them. If the different tests have been applied to different groups of individuals, samples are independent. In this case, XLSTAT uses a Student test to compare the AUCs (which requires assuming the normality of the AUC, which is acceptable if the samples are not too small). If different tests were applied to the same individuals, the samples are paired. In this case, XLSTAT calculates the covariance matrix of the AUCs as described by Delong and Delong (1988) based on Sen's work (1960), to then calculate the variance of the difference between two AUCs, and to calculate the p-value assuming the normality.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Event data: Select the data that correspond to the phenomenon being studied (for example, the presence or absence of a disease) and specify which code is associated to the **positive event** (for example D or + for a diseased individual).

Test data: Select the data that correspond to test value of the diagnostic. The data must be quantitative. If the data are ordinal, they must be recoded as quantitative data (for example 0,1,2,3,4). You must then specify if one should consider it as positive when the test value is greater or lower than a threshold value determined during the computations.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if, in column mode, the first row of the selected data contains a header, or in row mode, if the first column of the selected data contains a header.

Weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Options tab:

Confidence intervals:

- **Size (%)**: Enter the size of the confidence interval in % (default value: 95).
- **Wald**: Activate this option if you want to calculate confidence intervals on the various indexes using the approximation of the binomial distribution by the normal distribution. Activate "Adjusted" to use the adjustment of Agresti and Coull.
- **Wilson score**: Activate this option if you want to calculate confidence intervals on the various indexes using the Wilson score approximation.
- **Clopper-Pearson**: Activate this option if you want to calculate confidence intervals on the various indexes using the Clopper-Pearson approximation.
- **Continuity correction**: Activate this option if you want to apply the continuity correction to the Wilson score and to the interval on ratios.

A priori prevalence: If you know that the disease involves a certain proportion of individuals in the total population, you can use this information to adjust predictive values calculated from your sample.

Test on AUC: You can compare the AUC (Area Under the Curve) to 0.5, the value it would have if the test variable were purely random. This test is conducted using the method of calculating the variance chosen here above.

Costs: Activate this option if you want to evaluate the cost associated with the various possible decisions based on the threshold values of the test variable. You need to enter the costs that correspond to the different situations: TP (true positive), FP (false positive), FN (true negative), TN (true negative).

Data options tab:

Missing data:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Ignore missing data: Activate this option to ignore missing data.

Groups:

By group analysis: Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis on each group separately.

- **Compare**: Activate this option if want to compare the ROC curves, and perform the comparison tests.

Filter: Activate this option and select the data that describe to which group each observation belongs, if you want that XLSTAT performs the analysis for some groups that you will be able to select in a separate dialog box during the computations. If the "By group analysis" option is also activated, XLSTAT will perform the analysis for each group separately, only for the selected subset of groups.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

ROC analysis: Activate this option to display the table that lists the various indices calculated for each value of the test variable. You can choose to show or not show predictive values, likelihood ratios and of true/false positive and negative counts.

Test on the AUC: Activate this option if you want to display the results of the comparison of the AUC to 0.5, the value that corresponds to a random classifier.

Comparison of the AUCs: If you have selected several test variables or a group variable, activate this option to compare the AUCs obtained for the different variables or different groups.

Charts tab:

ROC curve: Activate this option to display the ROC curve.

True/False +/-: Activate this option to display the stacked bars chart that shows the % of the TP/TN/FP/FN for the different values of the test variable. The option is only available if the True/False +/- option in the Outputs tab is activated.

Decision plot: Activate this option to display the decision plot of your choice. This plot will help you to decide what level of the test variable is best.

Comparison of the ROC curves: Activate this option to display on a single plot the ROC curves that correspond to the various test variables or to the different groups. This option is only available if you select two or more test variables or if a group variable has been selected.

Results

Summary statistics: In this first table you can find statistics for the selected test(s), followed by a table recalling, for the phenomenon of interest, for the number of occurrences of each event and the prevalence of the positive event in the sample. The row displayed in bold corresponds to the positive event.

ROC curve: The ROC curve is then displayed. The strait dotted line that goes from (0 ;0) to (1 ;1) corresponds to the curve of a random test with no discrimination. The colored line corresponds to the ROC curve. Small squares correspond to observations (one square per observed value of the test variable).

ROC analysis: This table displays for each possible threshold value of the test variable, the various indices presented in the description section. On the line below the table you'll find a reminder of the rule set out in the dialog box to identify positive cases compared to the threshold value. Below the table you will find a stacked bars chart showing the evolution of the TP, TN, FP, FN depending on the value of the threshold value. If the corresponding option was activated, the **decision plot** is then displayed (for example, changes in the cost depending on the threshold value).

Area under the curve (AUC): This table displays the AUC, its standard error and a confidence interval.

Comparison of the AUC to 0.5: These results allow to compare the test to a random classifier. The confidence interval corresponds to the difference. Various statistics are then displayed including the p-value, followed by the interpretation of the comparison test.

Comparison of the AUCs: If you selected several test variables, once the above results are displayed for each variable, you will find the covariance matrix of the AUC, followed by the table of differences for each pair of AUCs with as comments the confidence interval, and then the table of the p-values. Values in bold correspond to significant differences. Last, a graph that compares the ROC curves displayed.

Example

An example showing how to compute ROC curves is available on XLSTAT Help Center:

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

An example showing how to compute ROC curves and compare them is available on XLSTAT Help Center:

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

References

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.

Parametric Illness-Death Model

Use this method to analyze survival-time data and model the movement of individuals among 3 states. This method could introduce covariates to estimate the effect of explanatory variables on the transition between states. It is available in Excel using the XLSTAT add-in statistical software.

In this section:

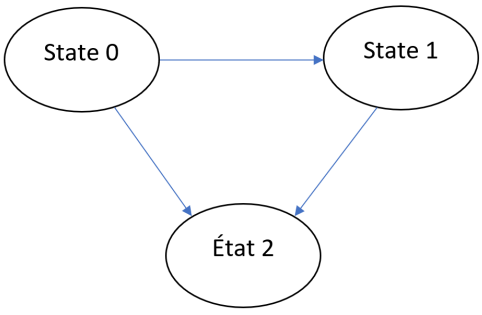
Description

Multi-state models are used when we observe more than 2 states. The illness-death models are a special case of multistate models with 3 states: the initial state, the transient state and the absorbing state — also called state 0, state 1 and state 2.

This model is frequently used in medical applications and research to analyze disease evolution or mortality. There are also many applications in the actuarial field to calculate the cost of insurance and viatical contracts.

Model

In XLSTAT, only irreversible Illness-Death models illustrated as followed can be fitted:



Let X be the stochastic process summarizing the state evolution of the individuals over time. Two functions define X: The transition probabilities and the transition intensities. In an Illness-Death model, we estimate the transition intensities, which are the analog of the risk function in survival analysis.

X is a non-homogeneous Markovian process, so the transition probabilities and intensities are expressed as follows from the state k to the state l:

- **Transition probabilities** p_{kl} only depend on the state in the present time s:

$$p_{kl}(s,t) = \mathbb{P}(X(t) = l | X(s) = k)$$

-**Transition intensities** $\alpha_{kl}(t)$ depend on the time t: $$ \alpha_{kl}(t)= \underset{\Delta t \to 0}{ lim } \frac{ p_{kl} (t,t+ \Delta t) }{ \Delta t } $$

The transition intensities are estimated by maximum likelihood, including censoring and truncated data.

Censoring and truncation

In survival analysis, truncation and censoring are phenomena that cause our samples to be incomplete. If we ignore them when analyzing our data, our estimates of population parameters will be inconsistent. The Illness-Death model in XLSTAT allows right censoring, interval censoring and left truncation, assuming that the censoring and truncation are non-informative.

Censoring occurs when we do not know the exact time of an event, while truncation occurs when we do not observe individuals with event times that are smaller or larger than certain values.

- Right censoring: The individual does not go to the absorbing state at the end of the observation. Given t the event time and t_{end} the end time observation, hence $t > t_{end}$
- Interval censoring: The event happend between two observations. Given t the event time and t_l and t_{l+1} two consecutive time observations, hence $t \in [t_l,\ t_{l+1}]$
- Left truncation: The event does not happen below a threshold. In our cases, the event cannot occur before the start of the study.

Parametric estimation of baseline intensities

In a parametric Illness-Death model, the baseline intensities are estimated by Weibull distributions. We denote a_{kl} and b_{kl} the form and rate parameters.

For the $k \to l$ transition, the baseline intensities $\alpha_{0,kl}$ can be expressed as follows:

$$ \alpha_{0,kl}(t) = a_{kl} \: \Big(\frac{1}{b_{kl}}\Big)^{a_{kl}} \: t^{a_{kl}-1}. $$

With the survival S and density f functions associated to the Weibull assumption:

$$ S(t) = e^{ - \Big(\dfrac{t}{b} \Big)^{a}} \qquad ;\qquad f(t) = a\: \Big(\dfrac{1}{b} \Big)^{a} \: t^{a-1} \: e^{ - \Big(\dfrac{t}{b} \Big)^{a}}. $$

The estimated parameters a_{kl} et b_{kl} are obtained by maximum likelihood.

Covariate effects

Transition intensities of an Illness-Death model consider the covariate effects. For example the effect of a treatment on a disease or on mortality.

Based on the Cox proportional hazards model, the intensities proportional hazards model expresses the transition intensities with the covariates. For the $k \to l$ transition, transition intensities has the form:

$$\alpha_{kl}(t) = \alpha_{0,kl}(t)\, e^{\beta_{kl}^T\, Z_{kl}}.$$

Here, $\alpha_{0,kl}$ is the baseline intensities (estimated by the Weibull distribution), β_{kl} is the effects or regression coefficients and Z_{kl} the covariate matrix.

This model involves 2 assumptions:

- The proportional hazards assumption suppose that the intensities ratio does not change over time:

$$\frac{\alpha(t|Z_i)}{\alpha(t|Z_j)} = \frac{\alpha_0(t) \times e^{\beta^T\, Z_i}}{\alpha_0(t) \times e^{\beta^T\, Z_j}} = e^{\beta^T (Z_i - Z_j)}.$$

However, this assumption has to be checked, and no simple method has been developed. To verify proportionality, the intensities at each time interval are estimated and checked to see whether the ratio is constant for all t.

To avoid this fastidious method, C. Touraine considers the age of the individual as a timescale to promote proportionality. Warning: This method does not guarantee proportionality.

- The log-linearity assumption, assumes that the log-scale transition intensities are linear:

$$log(\alpha(t|Z)) = log(\alpha_0(t)) + \beta_1\, Z_1 + ... + \beta_p\, Z_p.$$

This assumption has to be checked and no simple method has been developed. Considering binary covariates can avoid this assumption. Therefore, we recommend that the user include only binary covariates or quantitative covariates with tiny amplitude.

The results by covariates are calculated thanks to the formula of the proportional intensities model by setting the covariate Z_{kl} to the value of the modalities.

The β_{kl} are estimated by maximum likelihood.

Likelihood

Transition probabilities and likelihood of an irreversible Illness-Death model are calculated from the transition intensities α_{kl} and the survival function S_{kl}.

Censoring and truncation are included in the likelihood — that is why there are 7 different expressions of the likelihood. Each represents a combination of censoring and truncation. You can see these expressions in the works of C. Touraine (2013).

Since individuals do not have the same censoring and truncation type, the likelihood is calculated individually. The total likelihood denoted L is the product of the individual contribution L_i:

$$L = \prod_{i=1}^{n} L_i$$

To estimate the parameters we resort to numerical algorithms because the maximum of likelihood has no explicit solutions. Thus, the three baseline transition intensities and the regression parameters are estimated using the Levenberg-Marquardt's algorithm.

Levenberg-Marquardt Algorithm (LMA)

The Levenberg-Marquardt Algorithm (LMA) estimates transition intensities and regression parameters by maximizing the log-likelihood.

The LMA is a combination of two other methods:

- The Deepest Gradient: Maximizes the log-likelihood and updates the parameters at each iteration following the gradient direction.
- Newton-Raphson: Maximizes the log-likelihood derivatives and updated parameters at each iteration.

LMA alternates these two methods throughout the optimization. LMA acts like the Deepest Gradient method when the parameters are far from the solution; it acts like the Newton-Raphson method when the parameters are close to the solution. This makes the LMA faster than the gradient method (which suffers from various convergence problems) and more robust than the Newton method (which has a high computational cost).

Note: Despite the efficiency and robustness of the LMA, the Illness-Death model is a complex model and convergence is not guaranteed. Convergence failure may come from an insufficient number of observations, or from the data, for example, if one state is not represented in the dataset.

Predictions: Transition probabilities and life expectancy

Predictions give transition probabilities and life expectancies between 2 times:

- Entry time in the study
- End time in the study

7 transition probabilities are calculated:

- p_{00} the probability of remaining in the initial state
- p_{01} the probability of going from the initial state to the transient state
- p_{02} the probability of going from the initial state to the absorbing state
- p_{11} the probability of remaining in the transient state
- p_{12} the probability of going from the transient state to the absorbing state.

3 life expectancies are calculated:

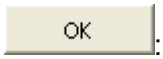
- E_{00} the life expectancy in state 0 (healthy life expectancy): Time expected for an individual to remain in the initial state given that he is in the initial state

- E_{02} the life expectancy: Time expected for an individual to stay alive (does not go in the absorbing state) given that he is in the initial state
- E_{12} the disease life expectancy: Time expected for an individual to stay alive (does not go in the absorbing state) given that he is in the transient state.

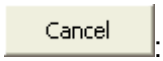
The formulas of these probabilities and life expectancies are given in Touraine C. (2013).

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Status indicator: Select here the 2 columns that correspond to the transient and absorbing state. All individuals are assumed to be in the initial state at the beginning of the study. These columns are binary, indicated by 1 if the individual was in the state and by 0 otherwise.

Time data: Select the time data corresponding to the 4 key ages: - Entry age: Age of the individual when they entered the study. - Left censoring age: Age of the individual... - before the transition to the transient state if the individual went there, - otherwise, at the last observation. - Right censoring age: Age of the individual... - at the time of the transition to the transient state if the individual went there, - otherwise, at the last observation. - Age of last news: Age of the individual... - at the time of the transition in the absorbing state if the individual went there - otherwise, at the last visit.

Range: If you choose this option, the results will be displayed from a cell in an existing sheet. You must select the cell first.

Sheet: Activate this option to display the results in a new sheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the selection contains the label of the variables.

Weights: Activate this option if you want to weigh the observations. Weights are considered to be 1 if this option is not activated. The weights must be positive integers. If the first row contains a label, make sure that the option "Variable labels" is activated.

Covariates tab:

Covariates: Activate this option if you want to add covariates to the model.

Different per transitions: Activate this option if you want to customize your covariates for each transition. If this option is not activated, the explanatory variables will be the same for all transitions.

Covariates: State 0 -> State 1 if "Different per transitions" is activated, **Covariates** otherwise: - **Quantitatives**: Activate this option if you want to add one or more quantitative covariates to the model. Then select the corresponding variable(s) on the Excel sheet. The selected data must be numeric. If the option **Different per transition** is activated, the selected data correspond to the transition "State 0 -> State 1", otherwise the selected data are applied to all transitions.

- **Qualitatives**: Activate this option if you want to add one or more qualitative covariates to the model. Then select the corresponding variable(s) on the Excel sheet. The selected data can be of any type, but numerical data are automatically considered nominal. If the option **Different per transition** is activated, the selected data correspond to the transition "State 0 -> State 1", otherwise the selected data are applied to all transitions.

If the variable labels have been selected, please check that the "Variable labels" option is activated.

If the option **"Different per transitions"** is activated:

Covariates: State 0 -> State 2": - **Quantitatives**: Activate this option if you want to add one or more quantitative covariates to the model for the transition State 0 -> State 2. Then select the corresponding variable(s) on the Excel sheet. The selected data must be numeric.

- **Qualitatives**: Activate this option if you want to add one or more qualitative covariates to the model for the transition State 0 -> State 2. Then select the corresponding variable(s) on the Excel sheet. The selected data can be of any type, but numerical data are automatically considered nominal.

If the variable labels have been selected, please check that the "Variable labels" option is activated.

Covariates: State 1 -> State 2": - **Quantitatives**: Activate this option if you want to add one or more quantitative covariates to the model for the transition State 1 -> State 2. Then select the corresponding variable(s) on the Excel sheet. The selected data must be numeric.

- **Qualitatives**: Activate this option if you want to add one or more qualitative covariates to the model for the transition State 1 -> State 2. Then select the corresponding variable(s) on the Excel sheet. The selected data can be of any type, but numerical data are automatically considered nominal.

If the variable labels have been selected, please check that the "Variable labels" option is activated.

Variable labels: Activate this option if the first row of the selected covariates contains the label of the covariates.

Options tab:

Significance level (%): Enter the value of the significance level to be used for the tests (default value: 5%). Confidence intervals will be also computed with this value.

The following options control the stopping conditions of the Levenberg-Marquardt algorithm (LMA):

- **Iterations**: Enter the maximum number of iterations for the LMA. The computations stop when the maximum number of iterations is reached. Default value: 100.
- **Convergence**: Enter the evaluation threshold value that considers that the algorithm has converged for:
    - **Parameters**: The sum of the square of the gradient (update of the parameters). Default value : 0.00001
    - **Likelihood**: The absolute difference of the log-likelihood between two iterations. Default value: 0.00001.
    - **Derivatives**: The calculation of the gradient and the hessian. Default value: 0.001.

Prediction tab:

General sub-tab:

Prediction: Activate this option if you want to make prediction. If the model contains covariates that were selected in the **Covariates** tab, covariates can be added to the prediction in the **Covariates** sub-tab (see below).

Entry time: Enter the age of the predicted individuals at the beginning of the study.

End time: Enter the age of the predicted individuals at the end of the study.

Covariates sub-tab:

Covariates: Activate this option if you want to add covariates to the predictions. If you choose this option, you must ensure that the prediction data are organized like the estimation data: Same variables and same order. This option is available only if covariates have been added to the model in the **Covariates** tab.

Variable labels: Activate this option if the first row of the selected covariates contains the label of the covariates.

Observation labels: Activate this option if you want to use observation labels available on the Excel sheet to display the results. The first row should not contain a header. If you do not activate this option, observation labels will be automatically created (PredObs1, PredObs2,...).

Missing data tab:

Do not accept missing data: Activate this option if you want XLSTAT to stop the calculation if missing values have been detected.

Remove the observation: Activate this option if you want to delete the observations with missing values.

Outputs tab:

Descriptive statistics: Activate this option to display the descriptive statistics for status indicator and time data.

Log-likelihood: Activate this option to display the value of the log-likelihood with and/or without covariates.

Weibull parameters: Activate this option to display the Weibull parameters for the 3 transitions.

Regression coefficients: Activate this option to display the table of regression coefficients. The first column indicates the transition and the second column the associated covariates. The third column gives the value of the coefficients. The fourth column shows the standard error of the coefficients, which measures the precision of the estimate. The fifth and sixth columns provide the result of the Wald test with the Wald statistic and its p-value. The last three columns show the Hazard Ratio with a confidence interval. This option is available only if covariates have been added to the model in the **Covariates** tab.

Survival distribution function: Activate this option to display the table of survival probabilities for each transition.

- **Confidence intervals**: Activate this option to display the confidence intervals associated to the table of survival probabilities.
- **Per covariates**: Activate this option to display the table of survival probabilities per covariates. This option is available only if covariates have been added to the model.

Transition probabilities: Activate this option to display the table of transition probabilities for each transition.

- **Confidence intervals**: Activate this option to display the confidence intervals associated to the table of transition probabilities.
- **Per covariates**: Activate this option to display the table of transition probabilities per covariates. This option is available only if covariates have been added to the model.

Transition intensities: Activate this option to display the table of transition intensities for each transition.

- **Confidence intervals**: Activate this option to display the confidence intervals associated to the table of transition intensities.
- **Per covariates**: Activate this option to display the table of transition intensities per covariates. This option is available only if covariates have been added to the model.

Charts tab:

Survival distribution function: Activate this option to display the chart of the survival distributions for each transition.

- **Confidence intervals**: Activate this option to display the confidence intervals associated to the chart of survival distributions.
- **Per covariates**: Activate this option to display the chart of survival distribution per covariates. This option is available only if covariates have been added to the model.

Transition probabilities: Activate this option to display the chart of transition probabilities for each transition.

- **Confidence intervals**: Activate this option to display the confidence intervals associated to the chart of transition probabilities.
- **Per covariates**: Activate this option to display the chart of transition probabilities per covariates. This option is available only if covariates have been added to the model.

Transition intensities: Activate this option to display the chart of transition intensities for each transition.

- **Confidence intervals**: Activate this option to display the confidence intervals associated to the chart of transition intensities.
- **Per covariates**: Activate this option to display the chart of transition intensities per covariates. This option is available only if covariates have been added to the model.

Results

XLSTAT offers a large number of tables and graphs to facilitate the analysis and interpretation of the results.

Descriptive statistics: A descriptive statistics table shows simple statistics for status indicator and time data. For time data, the number of observations, the amount of missing data, the amount of non-missing data, the mean and the standard deviation (unbiased) are displayed. For status indicator, the modalities, their numbers and their percentages are displayed.

Weibull parameters and **Regression coefficients**: These tables show the Weibull parameters and regression coefficients. Weibull parameters are used to calculate the baseline intensities. If covariates are included in the model, a regression coefficients table gives the coefficients estimation, the standard deviation, the χ^2-Wald and the p-value associated for each variable. And a Hazard Rate (exponential of the coefficients) is given with the associated confidence interval.

Transition intensities: This table shows the transition intensities, which is the analog of the risk function in survival analysis. The transition intensities help to compare the risk of transition between states.

Transition probabilities: This table shows the transition probabilities. These quantities help to compare the transition between states. More intuitive than the transition intensities, the transition probabilities help the user's interpretation.

Survival distribution: This table shows the survival distributions. These quantities help to compare the transition between states. More intuitive than the transition intensities, survival distributions help the user's interpretation.

Predictions: This table shows the predicted transition probabilities and the 3 life expectancies: In the common sense, of staying sick (remaining in the transient state) and staying healthy (remaining in the initial state). The predictions are computed between 2 times indicated by the user: Entry and end time of the study. If covariates have been added to the predictions, predictions are made for each individual with the covariates associated, otherwise the results correspond to a single individual.

Example

An example of the application of the parametric disease-death model is available on the XLSTAT Help Center at the following address:

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

References

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.

Laboratory data analysis

Method comparison

Use this tool to compare a method to a reference method or to a comparative method. Tests, confidence intervals are computed, and several plots are displayed to visualize differences, including the Bland Altman plot and the Difference plot. With this tool you are able to meet the recommendations of the Clinical and Laboratory Standards Institute (CLSI).

In this section:

Description

When developing a new method to measure the concentration or the quantity of an element (molecule, microorganism, …) you might want to check whether it gives results that are similar to a reference or comparative method or not. If there is a difference, you might be interested in knowing if this is due to a bias that depends on where you are on the scale variation. If a new measurement method is cheaper than a standard, but if there is a known and fixed bias, you might take into account the bias while reporting the results.

XLSTAT provides a series of tools to evaluate the performance of a method compared to another.

Repeatability analysis

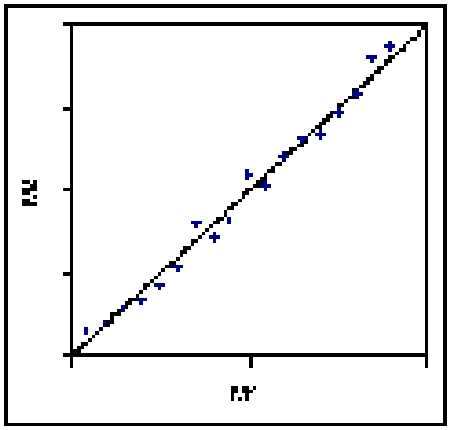
Repeatability and reproducibility analysis of measurement systems is available in the XLSTAT-SPC module (see gage R&R). The repeatability analysis provided here is a lighter version that is aimed at analyzing the repeatability of each method separately and to compare the repeatability of the methods. To evaluate the repeatability of a method, one needs to have several replicates. Replicates can be specified using the "Groups" field of the dialog box (replicates must have the same identifier). This corresponds to the case where several measures are taken on a given sample. If the method is repeatable, the variance within the replicates is low. XLSTAT computes the repeatability as a standard deviation and displays a confidence interval. Ideally, the confidence interval should contain 0.

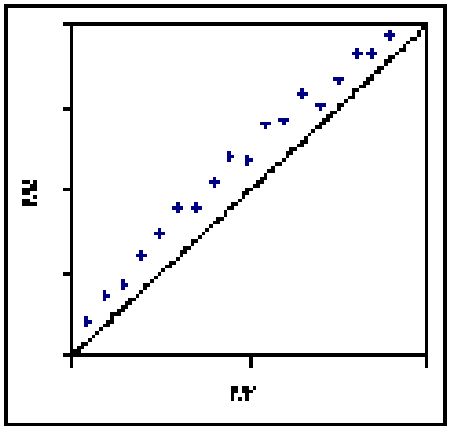
Paired t-test

Among the comparison methods, a paired t-test can be computed. The paired t-test allows to test the null hypothesis H0 that the mean of the differences between the results of the two methods is not different from 0, against an alternative hypothesis Ha that it is.

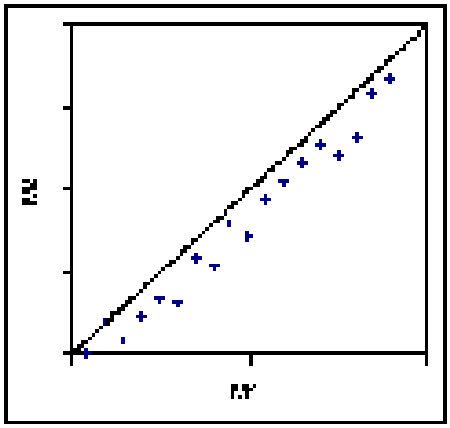
Scatter plots

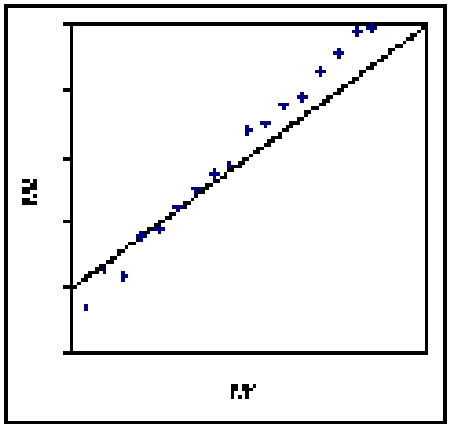
First, you can draw a scatter plot to compare the reference or comparative method against the method being tested. If the data are on both sides of the identity line (bisector) and close to it, the two methods give close and consistent results. If the data are above the identify line, the new method overestimates the value of interest. If the data are under the line, the new method underestimates the value of interest, at least compared to the comparative or reference method. If the data are crossing the identify line, there is a bias that depends on where you are on the scale of variation. If the data are randomly scattered around the identity line with some observations that are far from it, the new method is not performing well.



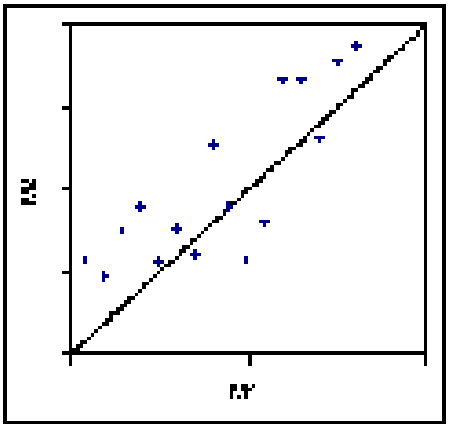


1. Consistent methods 2. Positive constant bias





3. Negative constant bias 4. Linear biais



5. Inconsistent methods

Bias

The bias is estimated as the mean of the differences between the two methods. If replicates are available, a first step computes the mean of the replicates. The standard deviation is computed as well as a confidence interval. Ideally, the confidence interval should contain 0.

Note: The bias is computed for the criterion that has been chosen for the Bland Altman analysis (difference, difference % or ratio).

Bland Altman and related comparison methods

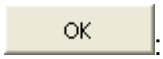
Bland and Altman recommend plotting the difference (T-S) between the test (T) and comparative or reference method (S) against the average (T+S)/2 of the results obtained from the two methods. In the ideal case, there should not be any correlation between the difference and the average whether there is a bias or not. XLSTAT tests whether the correlation is significantly different from 0 or not. Alternative possibilities are available for the ordinates of the plot: you can choose between the difference (T-S), the difference as a % of the sum (T-S)/(T+S), and the ratio (T/S). On the Bland Altman plot, XLSTAT displays the bias line, the confidence lines around the bias, and the confidence lines around the difference (or the difference % or the ratio).

Histogram and box plot

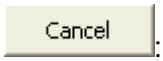
Histogram and box plots of the differences are plotted to validate the hypothesis that both are normally distributed, which is used to compute confidence intervals around the bias and the individual differences. When the size of the samples is small the histogram is of little interest and one should only consider the box plot. If the distribution does not seem to be normal, one might want to verify that point with a normality test, and one should consider with caution the confidence intervals.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Data (Method 1): Select the data that correspond to the first method, or to the reference method. If the name of the method is available in the first position of the data, make sure you activate the "Variable labels" option.

Data (Method 2): Select the data that correspond to the second method. If the name of the method is available in the first position of the data, make sure you activate the "Variable labels" option.

Groups: If replicates are available, select in this field the identifier of the measures. Two measures with the same group identifier are considered as replicates. XLSTAT uses the mean of the replicates for the analysis, and will provide you with repeatability results.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if, in column mode, the first row of the selected data contains a header, or in row mode, if the first column of the selected data contains a header.

Options tab:

Bland Altman analysis: Activate this option if you want to run a Bland Altman analysis and/or display a Bland Altman plot. Then, you need to specify the variable to use for the ordinates.

Difference analysis: Activate this option if you want to run a Difference analysis and/or display a Difference plot. Then, you need to specify the variable to use for the abscissa.

Significance level (%): Enter the size value of the significance level that is used to determine the critical value of the Student's t test and to generate the conclusion of the test.

Confidence intervals (%): Enter the size of the confidence interval in % (default value: 95).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Ignore missing data: Activate this option to ignore missing data. This option is only visible if the "Groups" option is active.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the two methods.

Paired t-test: Activate this option to display the results corresponding to a paired Student's t test to test whether the difference between the two methods is significant or not.

Bland Altman analysis: Activate this option to compute the Bias statistic and the corresponding confidence interval.

Charts tab:

Scatter plot: Activate this option to display the scatter plot showing on the abscissa the reference or comparative method, and on the ordinates the test method.

Bland Altman plot: Activate this option to display the Bland Altman plot.

Histogram: Activate this option to display the histogram of the differences (or differences % or ratios).

Box plot: Activate this option to display the box plot of the differences (or differences % or ratios).

Difference plot: Activate this option to display the difference plot.

Results

Summary statistics: In this first table you can find the basic descriptive statistics for each method.

t-test for two paired samples: These results correspond to the test of the null hypothesis that the two methods are not different versus the alternative hypothesis that they are. Note: this test is made using the assumption that the samples obtained with both methods are normally distributed.

A scatter plot is then displayed to allow comparing the two methods visually. The identity line is displayed on the plot. It corresponds to the ideal case where the samples on which the two methods are applied are identical and where the two methods would give exactly the same results.

The **Bland Altman analysis** is starts with an estimate of the bias, using the criterion that has been chosen (difference, difference in %, or ratio), the standard error and a confidence interval being as well displayed. The Bland Altman plot is displayed so that the difference between the two methods can be visualized. XLSTAT displays the correlation between the abscissa and the ordinates. One would expect it to be non-significantly different from 0, which means the confidence interval around the correlation should include 0.

The **histogram and the box plot** allow to visualize how the difference (or the difference % or the ratio) is distributed. A normality assumption is used when computing the confidence interval around the differences.

The **Difference plot** shows the difference between the two methods against the average of both methods, or against the reference method with an estimate of the bias, using the criterion that has been chosen (difference, difference in %, or ratio), the standard error and a confidence interval being as well displayed.

Example

An example showing how to compare two methods is available on XLSTAT Help Center:

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

References

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.

Passing and Bablok regression

Use this tool to compare two methods of measurement by a minimum of assumptions about their distribution.

In this section:

Description

Passing and Bablok (1983) developed a regression method that allows comparing two measurement methods (for example, two techniques for measuring concentration of an analyte), which overcomes the assumptions of the classical linear regression single that are inappropriate for this application. As a reminder the assumptions of the OLS regression are

- The explanatory variable, X in the model $y(i) = a + b \times x(i) + \varepsilon(i)$, is deterministic (no measurement error),
- The dependent variable Y follows a normal distribution with expectation aX
- The variance of the measurement error is constant.

Furthermore, extreme values can highly influence the model.

Passing and Bablok proposed a method which overcomes these assumptions: the two variables are assumed to have a random part (representing the measurement error and the distribution of the element being measured in medium) without needing to make assumption about their distribution, except that they both have the same distribution. We then define:

- $y(i) = a + b \times x(i) + \eta(i)$
- $x(i) = A + B \times y(i) + \xi(i)$

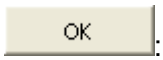
Where η and ξ follow the same distribution. The Passing and Bablok method allows calculating the a and b coefficients (from which we deduce A and B using $B = \frac{1}{b}$ and $A = -\frac{1}{b}$) as well as a confidence interval around these values. The study of these values helps comparing the methods. If they are very close, b is close to 1 and a is close to 0.

Passing and Bablok also suggested a linearity test to verify that the relationship between the two measurement methods is stable over the interval of interest. This test is based on a

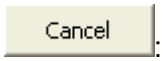
CUSUM statistic that follows a Kolmogorov distribution. XLSTAT provides the statistic, the critical value for the significance level chosen by the user, and the p-value associated with the statistic.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y method: Select the data that correspond to the method that will be displayed on the ordinates axis. If the name of the variable is available in the first position of the data, make sure you activate the "Variable labels" option.

X method: Select the data that correspond to the method that will be displayed on the abscissa axis. If the name of the variable is available in the first position of the data, make sure you activate the "Variable labels" option.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if, in column mode, the first row of the selected data contains a header, or in row mode, if the first column of the selected data contains a header.

Options tab:

Confidence intervals (%): Enter the size of the confidence interval in % (default value: 95).

Method:

- **Part I: same scale**: This method of estimation is the first method developed by Passing and Bablok (1983). It should be used when both methods are on the same scale and move in the same direction (positive correlation between X and Y).
- **Part III: different scale**: This method of estimation developed by Bablok *et al.* in 1988 is an improvement of the method known as *Part I*. It is more robust and can be used to compare two methods on different scales with possibly a negative correlation between X and Y.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the two methods.

Charts tab:

Predictions and residuals: Activate this option to display table corresponding to the input data, the predictions, the residuals and the perpendicular residuals.

Results

Summary statistics: In this first table you can find the basic descriptive statistics for each method.

Coefficients of the model: In this table are shown the coefficients a and b of the model and their respective confidence intervals.

Predictions and residuals: This table displays for each observation, the value of X, the value of Y, the model prediction, the residual and the perpendicular residual (the distance to the regression line by orthogonal projection) .

The charts allow to visualize the regression line, the observations and the model $Y = X$ (corresponding to the bisector of the plane) and the corresponding confidence interval calculated using the RMSE obtained from the model of Passing and Bablok but with the usual method for linear regression. This chart allows to visually check if the model is far from the model that would correspond to the hypothesis that the methods are identical.

Example

An example showing how to compare two methods using the Passing and Bablok regression is available on the XLSTAT Help Center:

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

References

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. *J. Clin. Chem. Clin. Biochem.* **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.

Deming regression

Use this tool to compare two methods of measurement with error on both measurements using Deming regression.

In this section:

Description

Deming (1943) developed a regression method, that allows comparing two measurement methods (for example, two techniques for measuring concentration of an analyte), which supposes that measurement error are present in both X and Y. It overcomes the assumptions of the classical linear regression that are inappropriate for this application. As a reminder the assumptions of the OLS regression are

- The explanatory variable, X in the model $y(i) = a + bx(i) + \epsilon(i)$, is deterministic (no measurement error),
- The dependent variable Y follows a normal distribution with expectation aX
- The variance of the measurement error is constant.

Furthermore, extreme values can highly influence the model.

Deming proposed a method which overcomes these assumptions: the two variables are assumed to have a random part (representing the measurement). The distribution has to be normal. We then define:

- $y(i) = y(i)^* + \epsilon(i)$
- $x(i) = x(i)^* + \eta(i)$

Assume that the available data $(y(i), x(i))$ are mismeasured observations of the "true" values $(y(i)^*, x(i)^*)$ where errors ϵ and η are independent. The ratio of their variances is assumed to be known:

$$\delta = \sigma^2(\eta)/\sigma^2(\epsilon)$$

XLSTAT-Life allows you to define variances of error measurement on X and Y.

We seek to find the line of "best fit" $y^* = a + bx^*$, such that the weighted sum of squared residuals of the model is minimized.

Where ϵ and η follow a normal distribution. The Deming method allows calculating the a and b coefficients as well as a confidence interval around these values. The study of these values helps comparing the methods. If they are very close, then b is close to 1 and a is close to 0.

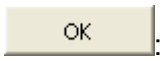
The Deming regression has two forms:

- Simple Deming regression: The error terms are constant. The estimation is very simple using a direct formula (Deming, 1943).
- Weighted Deming regression: The weighted Deming regression supposes that the error terms are not constant but only proportional. In that case, an iterative method is used to obtain parameters values (Linnet, 1990).

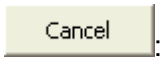
Confidence interval of the intercept and slope coefficient are complex to compute. XLSTAT-Life uses a jackknife approach to compute confidence intervals, as stated in Linnet (1993).

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

X: Select the data that correspond to the method that will be displayed on the abscissa axis. If the name of the variable is available in the first position of the data, make sure you activate the "Variable labels" option.

Y: Select the data that correspond to the method that will be displayed on the ordinates axis. If the name of the variable is available in the first position of the data, make sure you activate the "Variable labels" option.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if, in column mode, the first row of the selected data contains a header, or in row mode, if the first column of the selected data contains a header.

Constant error: Activate this option if the errors of both X and Y are supposed to be constant.

Proportional error: Activate this option if the errors of both X and Y are supposed to be proportional.

Options tab:

Confidence intervals (%): Enter the size of the confidence interval in % (default value: 95).

Variance of error on X: Enter the variance of the measurement error on X.

Variance of error on Y: Enter the variance of the measurement error on Y.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the two methods.

Charts tab:

Predictions and residuals: Activate this option to display table corresponding to the input data, the predictions and the residuals.

Results

Summary statistics: In this first table you can find the basic descriptive statistics for each method.

Coefficients of the model: In this table are shown the coefficients a and b of the model and their respective confidence intervals.

Predictions and residuals: This table displays for each observation, the value of X, the value of Y, the model prediction and the residuals.

The charts allow to visualize the regression line, the observations and the model $Y = X$ (corresponding to the bisector of the plane) and the corresponding confidence interval calculated using the RMSE obtained from the model of Deming but with the usual method for linear regression. This chart enables to visually check if the model is far from the model that would correspond to the hypothesis that the methods are identical.

Example

An example showing how to compare two methods using the Deming regression is available on the XLSTAT Help Center:

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

References

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.

Youden plots

Use this tool to create a Youden plot after collecting measurements to compare two materials A and B each evaluated by a series of laboratories.

In this section:

Description

Youden (1959) developed a procedure for representing data produced by N laboratories for two similar materials A and B (they can be identical when we want to compare measurement methods, or different but expected to give identical values). In his 1959 article, Youden insisted on the need for the method to be simple, so that the intervention of a statistical expert was not necessary. The objective here is to simply identify which laboratories are problematic, either because the two measurements performed show an abnormal difference, or because the two measurements are too different from what is obtained by other laboratories. Both inter-laboratory variability and intra-laboratory variability are analyzed here.

The result is a chart showing the measurements for material A on the abscissa axis and for material B on the ordinate axis. A circle, in the version described by Youden, is then displayed in order to be able to identify suspicious values, those which are outside the circle.

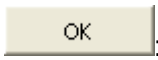
Originally, Youden made a prerequisite that the 2 materials tested are fairly close (two equivalent measurement techniques, two samples taken at two nearby locations) and that the measurements are on identical scales. However, if this is not the case, the XLSTAT user can request that the data be standardized. To standardize the two samples, the user has the option of choosing between a classic standardization (based on the arithmetic mean and unbiased variance), or a standardization based on robust statistics, as described in ISO 13528-2015-10 (algorithm A). Alternatively, the XLSTAT user can choose one of the two display methods which make it possible to get rid of data standardization. XLSTAT therefore offers three types of representation:

- Circle: Youden's procedure is directly applied.
- Ellipse: XLSTAT displays an ellipse around the observations. If the choice of robust statistics has been made, the robust covariance is used for the calculation of the ellipse (Maronna, 2019). The calculation of the ellipse can be done using a confidence interval based on either Fisher or the Chi-square distribution.

- Boxes: XLSTAT will display rectangles which on each axis surround the average by an interval of 2x2 and/or 2x3 times the measured standard deviation.

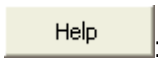
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.

: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Sample A: Select the data that correspond to the measurements obtained by each laboratory for sample A. If the name of the variable is available in the first position of the data, make sure you activate the "Variable labels" option.

Sample B: Select the data that correspond to the measurements obtained by each laboratory for sample B. If the name of the variable is available in the first position of the data, make sure you activate the "Variable labels" option.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if, in column mode, the first row of the selected data contains a header, or in row mode, if the first column of the selected data contains a header.

Observation labels: Check this option if you want to select the labels corresponding to each laboratory. If you do not check this option, labels will be created automatically (Obs1, Obs2, etc.). If a column header has been selected, check that the "Variable labels" option has been activated.

Weights: Activate this option if you want to identify laboratories which results are displayed on the plot, but that you do not want to included in the calculations of the statistics (mean and standard deviation). Simply put the value 0 for the weight to eliminate a laboratory from the calculations.

Options tab:

- **Circle**: Activate this option pour afficher un cercle.
- **Ellipse**: Activate this option to display an ellipse. You then have the choice between two approaches, based either on the Fisher or the Chi-square distribution.
- **Boxes**: Activate this option to display one or two confidence "boxes" based on twice and/or three times the standard deviation observed for each sample.

Confidence intervals (%): Enter the size of the confidence interval in % (default value: 95).

Standardize: Activate this option to standardize the measurements data.

Robust statistics: Activate this option to compute robust statistics for means, standard deviations and in the case where an ellipse is requested, for the covariance statistic.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the two methods.

Charts tab:

Predictions and residuals: Activate this option to display table corresponding to the input data, the predictions and the residuals.

Results

Summary statistics: In this first table you can find the basic descriptive statistics for each method.

Data and differences: In this table are displayed the data (standardized if the corresponding option has been activated) of samples A and B, as well as the differences and the absolute centered differences (the difference of the means is removed from the observed difference). The average of |D| is used for the calculation of Youden's circle.

Robust statistics: When requested, the robust statistics are displayed in this table.

Youden plot: This graph displays the data (standardized if the corresponding option has been activated). If the data is not standardized, a vertical line and a horizontal line passing through the means point are displayed. The straight line passing through the estimated means of the two samples (the origin if the data is standardized) and having an angle of 45° (circle option) or corresponding to the principal axis of the ellipse (ellipse option) is displayed. In case the boxes have been requested these are displayed on the graph.

Example

An example where measurements made by 29 laboratories for two samples are analyzed using a Youden plot is available on the XLSTAT Help Center:

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

References

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.

Dose effect analysis

Use this function to model the effects of a dose on a response variable, if necessary taking into account an effect of natural mortality.

In this section:

Description

This tool uses logistic regression (Logit, Probit, complementary Log-log, Gompertz models) to model the impact of doses of chemical components (for example a medicine or phytosanitary product) on a binary phenomenon (healing, death).

More information on logistic regression is available in the help section on this subject.

Natural mortality

This tool takes natural mortality into account in order to model the phenomenon studied more accurately. Indeed, if we consider an experiment carried out on insects, certain will die because of the dose injected, and others from other phenomenon. None of these associated phenomena are relevant to the experiment concerning the effects of the dose but may be taken into account. If p is the probability from a logistic regression model corresponding only to the effect of the dose, and if m is natural mortality, then the observed probability that the insect will succumb is:

$$P(obs) = m + (1 - m) \times p$$

Abbot's formula (Finney, 1971) is written as:

$$p = \frac{(P(obs) - m)}{(1 - m)}$$

The natural mortality m may be entered by the user as it is known from previous experiments, or is determined by XLSTAT.

ED50, ED90, ED99

XLSTAT calculates ED50 (or median dose), ED90 and ED99 doses which correspond to doses leading to an effect respectively on 50%, 90% and 99% of the population.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.

OK: Click this button to start the computations.

Cancel: Click this button to close the dialog box without doing any computation.

Help: Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Dependent variables:

Response variable(s): Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

Response type: Choose the type of response variable you have selected:

- **Binary variable**: If you select this option, you must select a variable containing exactly two distinct values. If the variable has value 0 and 1, XLSTAT will see to it that the high probabilities of the model correspond to category 1 and that the low probabilities correspond to category 0. If the variable has two values other than 0 or 1 (for example Yes/No), the lower probabilities correspond to the first category and the higher probabilities to the second.
- **Sum of binary variables**: If your response variable is a sum of binary variables, it must be of type numeric and contain the number of positive events (event 1) amongst those observed. The variable corresponding to the total number of events observed for this observation (events 1 and 0 combined) must then be selected in the "Observation

weights" field. This case corresponds, for example, to an experiment where a dose D (D is the explanatory variable) of a medicament is administered to 50 patients (50 is the value of the observation weights) and where it is observed that 40 get better under the effects of the dose (40 is the response variable).

Explanatory variables:

Quantitative: Activate this option if you want to include one or more quantitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The data selected may be of the numerical type. If the variable header has been selected, check that the "Variable labels" option has been activated.

Qualitative: Activate this option if you want to include one or more qualitative explanatory variables in the model. Then select the corresponding variables in the Excel worksheet. The selected data may be of any type, but numerical data will automatically be considered as nominal. If the variable header has been selected, check that the "Variable labels" option has been activated.

Model: Choose the type of function to use (see description).

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Observation weights: This field must be entered if the "sum of binary variables" option has been chosen. Otherwise, this field is not active. If a column header has been selected, check that the "Variable labels" option has been activated.

Options tab:

Firth's method: Activate this option to use Firth's penalized likelihood (see description).

Confidence interval (%): Enter the percentage range of the confidence interval to use for the various tests and for calculating the confidence intervals around the parameters and predictions. Default value: 95%.

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the Newton-Raphson algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 100.
- **Convergence**: Enter the maximum value of the evolution of the log of the likelihood from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.000001.

Take the log: Activate this option so that XLSTAT uses the logarithm of the input variables in the model.

Natural mortality parameter:

- **Optimised**: Choose this option so that XLSTAT optimizes the value of the natural mortality parameter.
- **User defined**: Choose this option to set the value of the natural mortality parameter.

Validation tab:

Validation: Activate this option if you want to use a sub-sample of the data to validate the model.

Validation set: Choose one of the following options to define how to obtain the observations used for the validation:

- **Random**: The observations are randomly selected. The "Number of observations" N must then be specified.
- **N last rows**: The N last observations are selected for the validation. The "Number of observations" N must then be specified.
- **N first rows**: The N first observations are selected for the validation. The "Number of observations" N must then be specified.
- **Group variable**: If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

Prediction tab:

Prediction: activate this option if you want to select data to use them in prediction mode. If activate this option, you need to make sure that the prediction dataset is structured as the

estimation dataset: same variables with the same order in the selections. On the other hand, variable labels must not be selected: the first row of the selections listed below must correspond to data.

Quantitative: activate this option to select the quantitative explanatory variables. The first row must not include variable labels.

Qualitative: activate this option to select the qualitative explanatory variables. The first row must not include variable labels.

Observations labels: activate this option if observations labels are available. Then select the corresponding data. If this option is not activated, the observations labels are automatically generated by XLSTAT (PredObs1, PredObs2 …).

Variable labels: Activate this option if the first row of the data selections (explanatory variables, observations labels) includes a header.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Correlations: Activate this option to display the explanatory variables correlation matrix.

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Type III analysis: Activate this option to display the type III analysis of variance table.

Model coefficients: Activate this option to display the table of coefficients for the model. Optionally, **confidence intervals** of type "*profile likelihood* " can be calculated (see description).

Standardized coefficients: Activate this option if you want the standardized coefficients (beta coefficients) for the model to be displayed.

Equation: Activate this option to display the equation for the model explicitly.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

Probability analysis: If only one explanatory variable has been selected, activate this option so that XLSTAT calculates the value of the explanatory variable corresponding to various probability levels.

Charts tab:

Regression charts: Activate this option to display regression chart:

- **Standardized coefficients**: Activate this option to display the standardized parameters for the model with their confidence interval on a chart.
- **Predictions**: Activate this option to display the regression curve.
- **Confidence intervals**: Activate this option to have confidence intervals displayed on charts (1) and (4).

Results

XLSTAT displays a large number tables and charts to help in analyzing and interpreting the results.

Summary statistics: This table displays descriptive statistics for all the variables selected. For the quantitative variables, the number of missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed. For qualitative variables, including the dependent variable, the categories with their respective frequencies and percentages are displayed.

Correlation matrix: This table displays the correlations between the explanatory variables.

Correspondence between the categories of the response variable and the probabilities: This table shows which categories of the dependent variable have been assigned probabilities 0 and 1.

Goodness of fit coefficients: This table displays a series of statistics for the independent model (corresponding to the case where the linear combination of explanatory variables reduces to a constant) and for the adjusted model.

- **Observations**: The total number of observations taken into account (sum of the weights of the observations);
- **Sum of weights**: The total number of observations taken into account (sum of the weights of the observations multiplied by the weights in the regression);
- **DF**: Degrees of freedom;

- **-2 Log(Like.)**: The logarithm of the likelihood function associated with the model;
- **R² (McFadden):** Coefficient, like the R^2, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model;
- **R²(Cox and Snell):** Coefficient, like the R^2, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to 1 minus the ratio of the likelihood of the adjusted model to the likelihood of the independent model raised to the power $\frac{2}{S_w}$, where S_w is the sum of weights.
- **R²(Nagelkerke):** Coefficient, like the R^2, between 0 and 1 which measures how well the model is adjusted. This coefficient is equal to ratio of the R^2 of Cox and Snell, divided by 1 minus the likelihood of the independent model raised to the power $\frac{2}{S_w}$;
- **AIC**: Akaike's Information Criterion;
- **SBC**: Schwarz's Bayesian Criterion.

Test of the null hypothesis H0: Y=p0: The H_0 hypothesis corresponds to the independent model which gives probability p_0 whatever the values of the explanatory variables. We seek to check if the adjusted model is significantly more powerful than this model. Three tests are available: the likelihood ratio test (-2 Log(Like.)), the Score test and the Wald test. The three statistics follow a χ^2 distribution whose degrees of freedom are shown.

Type III analysis: This table is only useful if there is more than one explanatory variable. Here, the adjusted model is tested against a test model where the variable in the row of the table in question has been removed. If the probability $Pr > LR$ is less than a significance threshold which has been set (typically 0.05), then the contribution of the variable to the adjustment of the model is significant. Otherwise, it can be removed from the model.

Model parameters: The parameter estimate, corresponding standard deviation, Wald's χ^2, the corresponding p-value and the confidence interval are displayed for the constant and each variable of the model. If the corresponding option has been activated, the "profile likelihood" intervals are also displayed.

The **equation of the model** is then displayed to make it easier to read or re-use the model.

The table of **standardized coefficients** (also called beta coefficients) are used to compare the relative weights of the variables. The higher the absolute value of a coefficient, the more important the weight of the corresponding variable. When the confidence interval around standardized coefficients has value 0 (this can be easily seen on the chart of normalized coefficients), the weight of a variable in the model is not significant.

The **predictions and residuals** table shows, for each observation, its weight, the value of the qualitative explanatory variable, if there is only one, the observed value of the dependent variable, the model's prediction, the same values divided by the weights, the standardized residuals and a confidence interval.

If only one quantitative variable has been selected, the **probability analysis** table allows to see to which value of the explanatory variable corresponds a given probability of success.

Example

A tutorial on how to use the dose effect analysis is available on the XLSTAT Help Center:

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

References

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 & Sons, New York.

Finney D.J. (1971). Probit Analysis. 3rd ed., Cambridge, London and New- York.

Firth D (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27-38.

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.

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.

Four/Five-parameter parallel lines logistic regression

Use this tool to analyze the effect of a quantitative variable on a response variable using the four/five-parameter logistic model. XLSTAT enables you to take into account some standard data while fitting the model, and to automatically remove outliers.

In this section:

Description

The four parameter logistic model writes:

$$y = a + \frac{d-a}{1+\left(\frac{x}{c}\right)^b} \tag{1.1}$$

where a, b, c, d are the parameters of the model, and where x corresponds to the explanatory variable and y to the response variable. a and d are parameters that respectively represent the lower and upper asymptotes, and b is the slope parameter. c is the abscissa of the mid-height point which ordinate is $(a+d)/2$. When a is lower than d, the curve decreases from d to a, and when a is greater than d, the curve increases from a to d.

The five parameter logistic model writes:

$$y = a + \frac{d-a}{\left[1+\left(\frac{x}{c}\right)^b\right]^e} \tag{1.2}$$

where the additional parameter e is the asymmetry factor.

The parallel lines four parameter logistic model writes:

$$y = a + \frac{d-a}{1+\left(s_0 \cdot \frac{x}{c_1} + s_1 \cdot \frac{x}{c_2}\right)^b} \tag{2.1}$$

where s_0 is 1 if the observation comes from the **standard sample**, and 0 if not, and where s_1 is 1 if the observation is from the **sample of interest**, and 0 if not. This is a constrained model because the observations corresponding to the standard sample influence the optimization of

the values of a, b, and d. From the above writing of the model, one can understand that this model generates two parallel curves, which only difference is the positioning of the curve, the shift being given by $(c_1 - c_0)$. If c_1 is greater than c_0, the curve corresponding to the sample of interest is shifted to the right of the curve corresponding to the standard sample, and vice-versa.

The parallel lines five parameter logistic model writes:

$$y = a + \frac{d - a}{[1 + (s_0 \cdot \frac{x}{c_0} + s_1 \cdot \frac{x}{c_1})^b]^e} \tag{2.2}$$

XLSTAT allows to fit:

- model 1.1 or 1.2 to a standard sample or to the sample of interest,
- model 2.1 or 2.2 to the standard sample and the sample of interest the same time.

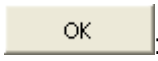
XLSTAT allows to either fit models 1.1 or 1.2 to a given sample (A case), or to fit models 1.1 or 1.2 to the standard (0) sample and then fit models 2.1 or 2.2 to both the standard sample and the sample of interest (B case).

If the Dixon's test option is activated, XLSTAT tests for each sample if some outliers influence too much the fit of the model. In the A case, a Dixon's test is performed once the model 1.1 or 1.2 is fitted. If an outlier is detected, it is removed, and the model is fitted again, and so on, until no outlier is detected. In the B case, we first perform a Dixon's test on the standard sample, then on the sample of interest, and then, the models 2.1 or 2.2 is fitted on the merged samples, without the outliers.

In the B case, and if the sum of the sample sizes is greater than 9, a Fisher's F test is performed to detect if the a, b, d and e parameters obtained with models 1.1 or 1.2 are not significantly different from those obtained with model 2.1 or 2.2.

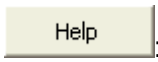
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.

: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y / Dependent variables:

Quantitative: Select the response variable(s) you want to model. If several variables have been selected, XLSTAT carries out calculations for each of the variables separately. If a column header has been selected, check that the "Variable labels" option has been activated.

X / Explanatory variables:

Quantitative: Select the quantitative explanatory variables to include in the model. If the variable header has been selected, check that the "Variable labels" option has been activated.

Model:

- **4PL**: Activate this option to fit the four parameter model.
- **5PL**: Activate this option to fit the five parameter model.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Subsamples: Activate this option then select a column (column mode) or a row (row mode) containing the sample identifier(s). The identifiers must be 0 for the standard sample and 1, 2, 3 ... for the other samples that you want to compare with the standard sample. If a header has been selected, check that the "Variable labels" option has been activated.

Options tab:

Initial values: Activate this option to give XLSTAT a starting point. Select the cells which correspond to the initial values of the parameters. The number of rows selected must be the same as the number of parameters.

Parameters bounds: Activate this option to give XLSTAT a possible region for all the parameters of the model selected. You must them select a two- column range, the one on the left being the lower bounds and the one on the right the upper bounds. The number of rows selected must be the same as the number of parameters.

Parameters labels: Activate this option if you want to specify the names of the parameters. XLSTAT will display the results using the selected labels instead of using generic labels pr1, pr2, etc. The number of rows selected must be the same as the number of parameters.

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 100.
- **Convergence**: Enter the maximum value of the evolution in the Sum of Squares of Errors (SSE) from one iteration to another which, when reached, means that the algorithm is considered to have converged. Default value: 0.00001.

Dixon's test: Activate this option to use the Dixon's test to remove outliers from the estimation sample.

Confidence intervals: Activate this option to enter the size of the confidence interval for the Dixon's test.

Missing data tab:

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Goodness of fit statistics: Activate this option to display the table of goodness of fit statistics for the model.

Model parameters: Activate this option to display the values of the parameters for the model after fitting.

Equation of the model: Activate this option to display the equation of the model once fitted.

Predictions and residuals: Activate this option to display the predictions and residuals for all the observations.

Charts tab:

Data and predictions: Activate this option to display the chart of observations and the curve for the fitted function.

- **Logarithmic scale**: Activate this option to use a logarithmic scale.

Residuals: Activate this option to display the residuals as a bar chart.

Residuals: Activate this option to display the chart of residuals versus the predictions.

Results

Summary statistics: This table displays for the selected variables, the number of observations, the number of missing values, the number of non- missing values, the mean and the standard deviation (unbiased).

If no group or a single sample was selected, the results are shown for the model and for this sample. If several sub- samples were defined (see sub-samples option in the dialog), the model is first adjusted to the standard sample, then each sub-sample is compared to the standard sample.

Fisher's test assessing parallelism between curves: The Fisher's F test is used to determine if one can consider that the models corresponding the standard sample and the sample of interest are significantly different or not. If the probability corresponding to the F value is lower than the significance level, then one can consider that the difference is significant.

Goodness of fit coefficients: This table shows the following statistics:

- The number of observations;
- The number of degrees of freedom (DF);
- The determination coefficient R^2;

- The sum of squares of the errors (or residuals) of the model (SSE or SSR respectively);
- The means of the squares of the errors (or residuals) of the model (MSE or MSR);
- The root mean squares of the errors (or residuals) of the model (RMSE or RMSR);

Model parameters: This table displays the estimator and the standard error of the estimator for each parameter of the model. It is followed by the **equation** of the model.

Predictions and residuals: This table displays giving for each observation the input data and corresponding prediction and residual. The outliers detected by the Dixon's test, if any, are displayed in bold.

Charts: On the first chart are displayed in blue color, the data and the curve corresponding to the standard sample, and in red color, the data and the curve corresponding to the sample of interest. A chart that allows to compare predictions and observed values as well as the bar chart of the residuals are also displayed. The last chart shows the residuals versus the predictions.

Example

A tutorial on how to use the four parameters logistic regression tool is available on the XLSTAT Help Center:

http://www.xlstat.com/demo-4pl.htm

References

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.

Differential expression

Use this tool to detect the most differentially expressed elements according to explanatory variables within a features/individuals data matrix that may be very large.

In this section:

Description

Differential expression lets you identify features (genes, proteins, metabolites…) that are significantly affected by explanatory variables. For example, we might be interested in identifying proteins that are expressed differentially between healthy and diseased individuals. In these kinds of studies, there is often a very large amount of data (= high-throughput data). At this stage, we may refer to 'omics' data analyses, in reference to analyses performed on the genome (genomics) or the transcriptome (transcriptomics) or the proteome (proteomics) or the metabolome (metabolomics), etc.

In order to test if features are differentially expressed, we often use traditional statistical tests. However, the size of the data may cause problems in terms of computation time as well as readability and statistical reliability of results. Those tools must therefore be slightly adapted in order to overcome these problems.

Empirical Bayes

This method is based on the adjustment of a linear model to each feature j. For this, the method of least squares is used to estimate all the coefficients $\hat{\alpha}_j$ such that $$ E(y_j) \approx X \tilde\alpha_j, \quad j = 1, n, $$

where X is the design matrix containing the values of a covariate for the different observations on the dependent variable. The variance of the k-th coefficient is supposed to check the relationship $$ var((\hat\alpha_j)*k) = \sigma^2_j((X^T X)^{-1}k}) $$

By replacing $sigma_j^2$ with an estimate, it is then possible to perform various statistical tests on these coefficients. In the specific case of the Empirical Bayes test, the moderate t-statistic is defined as: $$ t = \frac{\hat\beta_j}{\tilde s_j \sqrt{v_j}} $$

where $\hat{\beta}_j = C\hat{\alpha}_j$ is a linear combination of the estimated coefficients, $v_j = C^T (X^T X)^{-1} C$ and $\tilde{s}$ is the posterior estimate of the variance. One assumes that the real variance of the

feature j, if not differentially expressed, is taken from the distribution: $$ \frac{1}{\sigma_j^2} \approx \frac{1}{d_0 s_0^{2}\chi^2_{d_0} $$

where s_0 is the variance expected value. From this last equation, it can be shown that $$ \tilde s_j^2 = \frac{d_0 s_0^2 + d_j s_j^2}{d_0+d_j}. $$

Statistical tests

Statistical tests proposed within the differential expression tool are traditional parametric or non-parametric tests: Student t-test, ANOVA, Mann-Whitney, Kruskal-Wallis).

Statistical tests proposed within the differential expression tool are basic one, parametric or non-parametric tests, and documented in other sections of the help: Student t-test, ANOVA, Mann-Whitney, Kruskal-Wallis. A third, less conventional, test is available, which is based on the empirical Bayes approach described below.

Post-hoc corrections

The p-value represents the risk that we will be wrong when stating that an effect is statistically significant. Running a test several times increases the number of computed p-values, and subsequently, the risk of detecting significant effects which are not significant in reality. Considering a significance level alpha of 5%, we would likely find 5 significant p-values by chance over 100 computed p-values. When working with high-throughput data, we often test the effect of an explanatory variable on the expression of thousands of genes, thus generating thousands of p-values. Consequently, p-values should be corrected (= increased = penalized) as their numbers grow. XLSTAT proposes three common p-value correction methods:

Benjamini-Hochberg: this procedure makes sure that p-values increase both with their number and the proportion of non-significant p-values. It is part of the FDR (False Discovery Rate) correction procedure family. The Benjamini-Hochberg correction is poorly conservative (= not very severe). It is therefore adapted to situations where we are looking for a large number of genes that are likely affected by the explanatory variables. It is widely used in differential expression studies.

The corrected p-value according to the Benjamini-Hochberg procedure is defined by:

$$p_{BenjaminiHochberg} = min(p \times nbp/j, 1)$$

where p is the original (uncorrected) p-value, nbp is the number of computed p-values in total, and j is the rank of the original p-value when p-values are sorted in ascending order.

Benjamini-Yekutieli: this procedure makes sure that p-values increase both with their number and the proportion of non-significant p-values. It is part of the FDR (False Discovery Rate) correction procedure family. In addition to Benjamini-Hochberg's approach, it takes into account a possible dependence between the tested features, making it more conservative than this procedure. However, it is far less stringent than the Bonferroni approach, which we describe just after.

The corrected p-value according to the Benjamini-Yekutieli procedure is defined by:

$$p_{BenjaminiYekutieli} = min[(p \times nbp \sum_{i=1}^{nbp} 1/i)/j, 1]$$

where p is the original p-value, nbp is the number of computed p-values in total and j is the rank of the original p-value when p-values are sorted in ascending order.

Bonferroni: p-values increase only with their number. This procedure is very conservative. It is part of the FWER (Familywise error rate) correction procedure family. It is rarely used in differential expression analyses. It is useful when the goal of the study is to select a very low number of differentially expressed features.

The corrected p-value according to the Bonferroni procedure is defined by:

$$p_{Bonferroni} = min(p \times nbp, 1)$$

where p is the original p-value and nbp is the number of computed p-values in total.

Multiple pairwise comparisons

After one-way ANOVAs or Kruskal-Wallis tests, it is possible to perform multiple pairwise comparisons for each feature taken separately. XLSTAT provides different options including:

- Tukey's HSD test: this test is the most used (HSD: Honestly Significant Difference).
- Fisher's LSD test: this is a Student's test that tests the hypothesis that all the means for the various categories are equal (LSD: Least Significant Difference).
- Bonferroni's t* test: this test is derived from the Student's test and is less reliable as it takes into account the fact that several comparisons are carried out simultaneously. Consequently, the significance level of the test is modified according to the following formula:

$$\alpha^{'} = \alpha/(g(g-1)/2)$$

where g is the number of categories of the factor whose categories are being compared.

- Dunn-Sidak's test: This test is derived from Bonferroni's test. It is more reliable in some situations.

$$\alpha' = 1 - (1-\alpha)^{2/[g(g-1)]}$$

Non-specific filtering

Before launching the analyses, it is a good idea to filter out features with very poor variability across individuals. Non-specific filtering has two major advantages:

- It allows computations to focus less on features that are very likely to be not differentially expressed, thus saving computation time.

- It limits post-hoc penalizations, as fewer p-values are computed.

Two methods are available in XLSTAT:

- The user specifies a variability threshold (interquartile range or standard deviation), and features with lower variability are eliminated prior to analyses.

- The user specifies a percentage of features with low variability (interquartile range or standard deviation) to be removed prior to analyses.

Biological effects and statistical effects: the volcano plot

A statistically significant effect is not necessarily interesting at the biological scale. An experiment involving very precise measurements with a high number of replicates may provide low p-values associated with very weak biological differences. It is thus a good idea to keep an eye on biological effects and not to rely solely on p-values. The volcano plot is a scatter chart that combines statistical effects on the y-axis and biological effects on the x-axis for a whole features/individuals data matrix. The only constraint is that it can only be executed to examine the difference between the levels of two-level qualitative explanatory variables.

The y axis coordinates are -log10(p-values), making the chart easier to read: high values reflect the most significant effects, whereas low values correspond to effects which are less significant.

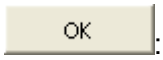
XLSTAT provides two ways of building the x axis coordinates:

- Difference between the mean of the first level and the mean of the second for each feature. Generally, we use this format when handling data on a transformed scale such as log or square root.

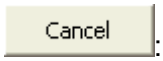
- Log2 of fold change between the two means: log2(mean1/mean2). This format should preferably be used with untransformed data.

Dialog box

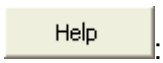
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Features/individuals table: Select the features/individuals data matrix in the Excel worksheet. The data selected must be numeric.

Data format:

Features in rows: Activate this option if features are stored in lines and individuals (or samples) are stored in columns.

Features in columns: Activate this option if features are stored in columns and individuals (or samples) are stored in lines.

X / Explanatory variables: Select one or more qualitative explanatory variables. The selected data can be of any type, but the numerical data are automatically considered as nominal. If variable labels have been selected, please verify that the “Variable labels” option is enabled.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet in the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated, you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Options tab:

Test type:

Parametric: Activate this option if you want to run parametric tests (Student or one-factor ANOVA)

Nonparametric: Activate this option if you want to use the Kruskal-Wallis nonparametric test. This option can only be activated if a single explanatory variable is selected, and this variable is qualitative.

Empirical Bayes: Activate this option if you want to run the empirical Bayesian test.

Level of significance (%): Enter the level of significance to be used for the different tests (default value: 5%).

Post-hoc corrections: Select the type of correction of the p-values desired (Benjamini-Hochberg, Benjamini-Yekutieli, Bonferroni, no correction; see Description).

p-values to keep: Enter the number of lowest p-values to display in the results. If the number entered is greater than the number of characters in the filtered character/person array, XLSTAT will display the p-values associated with all characters. Default value: 100.

Multiple pairwise comparisons: Activate this option and choose the comparison method if you wish. Information about the multiple comparisons tests are available in the description section.

Bonferroni correction: Activate this option if you wish penalize multiple pairwise corrections using the Bonferroni method.

Non-specific filtering: Activate this option to filter out features with low variability prior to computations.

Criterion and **threshold**: Select the non-specific filtering criterion.

- **Standard deviation<**: all features with a standard deviation lower than the selected threshold are removed.
- **Interquartile range<**: all features with an interquartile range lower than the selected threshold are removed.
- **%(Std. dev.)**: a percentage of features with low standard deviation are removed. The percentage should be indicated in the threshold box
- **%(IQR)**: a percentage of features with low interquartile range is removed. The percentage should be indicated in the threshold box.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

Charts tab:

Color scale: select the color range to use in the heat map (red to green through black; red to blue through white; red to yellow).

Width and **height**: select a magnification factor for the heat map's width or height.

Results

Summary statistics: The tables of descriptive statistics show the simple statistics for all individuals. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed.

Heat map: The features dendrogram is displayed vertically (rows) and the individuals dendrogram is displayed horizontally (columns). A heat map is added to the chart, reflecting data values.

Similarly expressed features are characterized by horizontal rectangles of homogeneous color along the map.

Similar individuals are characterized by vertical rectangles of homogeneous color along the map.

Clusters of similar individuals characterized by clusters of similarly expressed features can be detected by examining rectangles or squares of homogeneous color at the intersection between feature clusters and individual clusters inside the map.

Example

An example showing how to compare two methods using the Deming regression is available on the XLSTAT Help Center:

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

References

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

Use this tool to perform clustering on both columns and rows of a features/individuals data matrix, and to draw heat maps.

In this section:

Description

While exploring features/individuals matrices, it is interesting to examine how correlated features (i.e. genes, proteins, metabolites) correspond to similar individuals (i.e. samples). For example, a cluster of diseased kidney tissue samples may be characterized by a high expression of a group of genes, compared to other samples. The heat maps tool in XLSTAT allows performing such explorations.

How it works in XLSTAT

Both features and individuals are clustered independently using ascendant hierarchical clustering based on Euclidian distances, optionally preceded by the k-means algorithm depending on the matrix's size. The data matrix's rows and columns are then permuted according to corresponding clusterings, which brings similar columns closer to each other and similar lines closer to each other. A heat map is then displayed, reflecting data in the permuted matrix (data values are replaced by corresponding color intensities).

Non-specific filtering

Before launching the analyses, it is interesting to filter out features with very poor variability across individuals. In heat map analysis, non-specific filtering has two major advantages:

- It allows computations to focus less on features which are very likely to be not differentially expressed thus saving computation time.

- It improves the readability of the heat map chart.

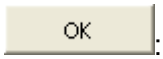
Two methods are available in XLSTAT:

- The user specifies a variability threshold (interquartile range or standard deviation), and features with lower variability are eliminated prior to analyses.

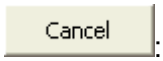
- The user specifies a percentage of features with low variability (interquartile range or standard deviation) to be removed prior to analyses.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Features/individuals table: Select the features/individuals data matrix in the Excel worksheet. The data selected must be of type numeric.

Data format:

Features in rows: activate this option if features are stored in lines and individuals (or samples) are stored in columns.

Features in columns: activate this option if features are stored in columns and individuals (or samples) are stored in lines.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if feature and individual labels are included in the selection.

Cluster features: Activate this option if you wish the heat map to include clustering on features

Cluster individuals: Activate this option if you wish the heat map to include clustering on individuals (or samples).

Options tab:

Center: Activate this option to center each row separately.

Reduce: Activate this option to reduce each row separately.

Non-specific filtering: Activate this option to filter out features with low variability prior to computations.

Criterion and **threshold**: Select the non-specific filtering criterion.

- **Standard deviation<**: all features with a standard deviation lower than the selected threshold are removed.
- **Interquartile range<**: all features with an interquartile range lower than the selected threshold are removed.
- **%(Std. dev.)**: a percentage of features with low standard deviation are removed. The percentage should be indicated in the threshold box
- **%(IQR)**: a percentage of features with low interquartile range are removed. The percentage should be indicated in the threshold box.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- Mean or mode: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- Nearest neighbor: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

Charts tab:

Color scale : select the color range to use in the heat map (red to green through black; red to blue through white; red to yellow).

Width and **height** : select a magnification factor for the heat map's width or height.

Color calibration:

- **Automatic**: Activate this option if you want XLSTAT to automatically choose boundary values that will delimit the heat map color range.
- **User defined**: Activate this option if you want to manually choose the minimum (Min) and maximum (Max) values that will delimit the heat map color range.

Results

Summary statistics: The tables of descriptive statistics show the simple statistics for all individuals. The number of observations, missing values, the number of non-missing values, the mean and the standard deviation (unbiased) are displayed.

Heat map: The features dendrogram is displayed vertically (rows) and the individuals dendrogram is displayed horizontally (columns). A heat map is added to the chart, reflecting data values.

Similarly expressed features are characterized by horizontal rectangles of homogeneous color along the map.

Similar individuals are characterized by vertical rectangles of homogeneous color along the map.

Clusters of similar individuals characterized by clusters of similarly expressed features can be detected by examining rectangles or squares of homogeneous color at the intersection between feature clusters and individual clusters inside the map.

Example

A tutorial on two-way clustering is available on the XLSTAT Help Center:

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

References

Hahne F., Huber W., Gentleman R. and Falcon S. (2008). Bioconductor Case Studies. Springer.

Inter-laboratory proficiency testing

Use this tool to perform proficiency testing on one or more participants (laboratories, inspection bodies, or individuals), where one ore more tests or measurements (tests in XLSTAT) have been recorded for each of them.

In this section:

Description

Proficiency testing involves using statistical methods to compare the performance of several participants (which may be laboratories, inspection bodies, or individuals), referred to as “participants” in XLSTAT, for specific measurements (referred to as “tests” in XLSTAT). Proficiency testing can be performed to assess the performance of laboratories making measurements, to detect problems in one or more laboratories when they arise, or to establish effectiveness and comparability of different methods.

The methods consist of identifying, then removing or ignoring outliers and producing robust estimates of both location and scale estimators.

This tool is based on the ISO-13528 standard. It was first developed using the 2015 edition. As some errors have been detected in the 2015 version of the document (see Fahmy, 2021 for further details), XLSTAT also lets users run the analysis including the errors.

XLSTAT allows for a highly automated analysis of the laboratory data and can yield a series of general and robust statistics that can then be used to interpret the results and define to what extent the proficiency standards are met. While many of the XLSTAT functions can be used to analyze inter-laboratory data, the automation provided here is very useful when the user wants to use recommended but more complex algorithms, such as Algorithm A, Algorithm S or the Q/Hampel approach.

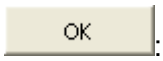
The Circle Technique (Van Nuland, 1992) is a very effective and underused technique to simultaneously compare the means and the variances of several participants, while allowing you to identify potential outliers.

This tool is still evolving and you are welcome to submit any feedback or request.

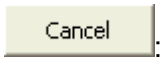
Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various

elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Participants/Tests or Tests/Participants table: Select the data corresponding to the data collected for each item with one or more tests being recorded for each item (typically participants are laboratories and tests are measurements). You may include in the selection the participants and tests labels. In that case, check the corresponding options.

Data format: * **Participants/Tests table**: Choose this option if rows correspond to participants and columns to tests. * **Tests/Participants table**: Choose this option if rows correspond to tests and columns to participants.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Participants labels: Activate this option if the participants labels are included in the selection.

Tests labels: Activate this option if the tests labels are included in the selection.

Options tab:

Location: Choose whether you want to use the mean or the median as the location statistic.

ISO-13528-2015 errors: Activate this option if you want XLSTAT to mimic the errors in the computation of the Q_n statistic.

Algorithm S:

- **Scale**: Choose whether you want to use the range or the standard deviation as the scale statistic.

Algorithm A:

- **Scale**: Choose whether you want to use the range, the standard deviation using the Grubbs approach to remove outliers, the nIQR, Q_n or Q as the scale statistic.
- **Only if MAD=0**: Activate this option to replace the MAD with the median of the absolute differences with the mean, if the MAD=0.
- **Update $s^\star$**: Activate this option if you want to update the robust estimate $s^\star$ at each iteration.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Ignore missing data: Activate this option so that missing data are ignored (removed only when necessary).

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics.

Z-scores: Activate this option to display the Z-scores. Several methods are available to compute Z-scores: * **Mean/Std. deviation**: This is the usual way to compute Z-scores. The centering is done using the mean, and the rescaling is done using the standard deviation. * $\mathbf{m^\star/s^\star}$: The centering and rescaling is done using the robust statistics $m^\star$ and $s^\star$. Note that the formula for the denominator is $scale = 1.25\frac{s^\star}{\sqrt{r}}$ where r is the number of repetitions. * **Reference/Std. deviation**: The centering uses a reference value entered by the user, and the rescaling is done using the standard deviation. * **Reference/$\mathbf{s^\star}$**: The centering uses a reference value entered by the user, and the rescaling is done using the robust statistic $s^\star$. Note that the formula for the denominator is $scale = 1.25\frac{s^\star}{\sqrt{r}}$ where r is the number of repetitions.

Charts tab tab:

Homescedasticity plot: Activate this option to display the "circle plot" with means in abscissa and standard deviations on ordinates. * **Participants labels** : Activate this option to label the points using the participants names. * For the scale statistic, two options are available. You can select to plot standard deviations or ranges againts the means.

Z-scores control chart: Activate this option to display the control chart of the Z-scores. The control limits at [-2,2] and [-3,3] are displayed.

Results

XLSTAT displays several tables and a chart to help in analyzing and interpreting the results.

If there are multiple participants (laboratories) and if two or more tests (measurements) have been performed for each of them, XLSTAT displays a list of summary statistics for each of them, including robust statistics.

Summary statistics are then computed across all participants.

If possible, the homoscedasticity plot is displayed to compare the location and scale of the different participants using the Circle Technique. Confidence lines (90%, 95%, 99%) are displayed to identify participants that show values that are potential outliers.

Last, if there are multiple participants (laboratories) and if two or more tests (measurements) have been performed, XLSTAT displays the results of the advanced algorithms described in ISO-13528 (Algorithm A, Algorithm S, Q/Hampel) that are designed to compute location and scale estimators iteratively. Algorithm A and the Q/Hampel method are used to obtain robust estimators of location and scale. Algorithm S is used to estimated the scale estimator from standard deviations or ranges.

Example

A tutorial on how to use inter-laboratory proficiency testing is available on the XLSTAT Help Center:

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

References

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.

Multiblock analysis

Canonical Correlation Analysis (CCorA)

Use Canonical Correlation Analysis (CCorA, sometimes CCA, but we prefer to use CCA for Canonical Correspondence Analysis) to study the correlation between two sets of variables and to extract from these tables a set of canonical variables that are as much as possible correlated with both tables and orthogonal to each other.

In this section:

Description

Canonical Correlation Analysis (CCorA, sometimes CCA, but we prefer to use CCA for Canonical Correspondence Analysis) is one of the many methods that allow to study the relationship between two sets of variables. Discovered by Hotelling (1936) this method is used a lot in ecology but is has been supplanted by RDA (Redundancy Analysis) and by CCA (Canonical Correspondence Analysis).

This method is symmetrical, contrary to RDA, and is not oriented towards prediction. Let Y_1 and Y_2 be two tables, with respectively p and q variables. CCorA aims at obtaining two vectors a_i and b_i such that

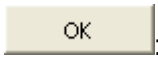
$$\rho(i) = \frac{corr(Y_1 a_i, Y_2 b_i)}{var(Y_1 a_i) var(Y_2 b_i)}$$

is maximized. Constraints must be introduced so that the solution for a_i et b_i is unique. As we are in the end trying to maximize the covariance between $Y_1 a_i$ and $Y_2 b_i$ and to minimize their respective variance, we might obtain components that are well correlated among each other, but that are not explaining well Y_1 and Y_2. Once the solution has been obtained for i=1, we look for the solution for i=2 where a_2 and b_2 must respectively be orthogonal to a_1 and b_1, and so on. The number of vectors that can be extracted is to the maximum equal to $min(p, q)$.

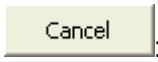
Note: The inter-batteries analysis of Tucker (1958) is an alternative where one wants to maximize the covariance between the $Y_1 a_i$ and $Y_2 b_i$ components.

Dialog box

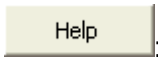
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Y1: Select the data that corresponds to the first table. If the "Column labels" option is activated (column mode) you need to include a header on the first row of the selection. If the "Row labels" option is activated (row mode) you need to include a header in the first column of the selection in the selection.

Y2: Select the data that corresponds to the second table. If the "Column labels" option is activated (column mode) you need to include a header on the first row of the selection. If the "Row labels" option is activated (row mode) you need to include a header in the first column of the selection in the selection.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column/Row labels: Activate this option if, in column mode, the first row of the selected data contains a header, or in row mode, if the first column of the selected data contains a header.

Observation labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Column labels" option is activated you need to include a header in the selection. If this option is not activated, the sites labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Options tab:

Type of analysis: Select from which type of matrix the canonical analysis should be performed.

Y1:

- **Center**: Activate this option to center the variables of table Y1.
- **Reduce**: Activate this option to standardize the variables of table Y1.

Y2:

- **Center**: Activate this option to center the variables of table Y2.
- **Reduce**: Activate this option to standardize the variables of table Y2.

Filter factors: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %**: Activate this option then enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Activate this option to set the number of factors to take into account.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

Covariance/Correlations /[Y1Y2]'[Y1Y2]: Activate this option to display the similarity matrix that is being used.

Eigenvalues: Activate this option to display the table and the scree plot of the eigenvalues.

Wilks Lambda test: Activate this option to display the results of the Wilks lambda test.

Canonical correlations: Activate this option to display the canonical correlations.

Redundancy coefficients: Activate this option to display the redundancy coefficients.

Canonical coefficients: Activate this option display the canonical coefficients.

Variables/Factors correlations: Activate this option to display the correlations between the initial variables of Y1 and Y2 with the canonical variables.

Canonical variables adequacy coefficients: Activate this option to display canonical variables adequacy coefficients.

Squared cosines: Activate this option to display the squared cosines of the initial variables in the canonical space.

Scores: Activate this option to display the coordinates of the observations in the space of the canonical variables.

Charts tab:

Correlation charts: Activate this option to display the charts involving correlations between the components and the variables.

- **Vectors**: Activate this option to display the variables with vectors.
- **Colored labels**: Activate this option to display the labels with the same color as the corresponding points. If this option is not activated the labels are displayed in black.

Results

Summary statistics: This table displays the descriptive statistics for the objects and the explanatory variables.

Similarity matrix: The matrix that corresponds to the "type of analysis" chosen in the dialog box is displayed.

Eigenvalues and percentages of inertia: In this table are displayed the eigenvalues, the corresponding inertia, and the corresponding percentages. Note: in some software, the

eigenvalues that are displayed are equal to L / (1-L), where L is the eigenvalues given by XLSTAT.

Wilks Lambda test: This test allows to determine if the two tables Y1 and Y2 are significantly related to each canonical variable.

Canonical correlations: The canonical correlations, bounded by 0 and 1, are higher when the correlation between Y1 and Y2 is high. However, they do not tell to what extent the canonical variables are related to Y1 and Y2. The squared canonical correlations are equal to the eigenvalues and, as a matter of fact, correspond to the percentage of variability carried by the canonical variable.

The results listed below are computed separately for each of the two groups of input variables.

Redundancy coefficients: These coefficients allow to measure for each set of input variables what proportion of the variability of the input variables is predicted by the canonical variables.

Canonical coefficients: These coefficients (also called *Canonical weights*, or *Canonical function coefficients*) indicate how the canonical variables were constructed, as they correspond to the coefficients in the linear combine that generates the canonical variables from the input variables. They are standardized if the input variables have been standardized. In that case, the relative weights of the input variables can be compared.

Correlations between input variables and canonical variables (also called *Structure correlation coefficients,* or *Canonical factor loadings*) allow understanding how the canonical variables are related to the input variables.

The **canonical variable adequacy coefficients** correspond, for a given canonical variable, to the sum of the squared correlations between the input variables and canonical variables, divided by the number of input variables. They give the percentage of variability taken into account by the canonical variable of interest.

Square cosines: The squared cosines of the input variables in the space of canonical variables allow to know if an input variable is well represented in the space of the canonical variables. The squared cosines for a given input variable sum to 1. The sum over a reduced number of canonical axes gives the communality.

Scores: The scores correspond to the coordinates of the observations in the space of the canonical variables.

Example

An example of Canonical Correlation Analysis is available on XLSTAT Help Center:

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

References

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.

Redundancy Analysis (RDA)

Use Redundancy Analysis (RDA) to analyze a table of response variables using the information provided by a set of explanatory variables, and visualize on the same plot the two sets of variables, and the observations.

In this section:

Description

Redundancy Analysis (RDA) was developed by Van den Wollenberg (1977) as an alternative to Canonical Correlation Analysis (CCorA). RDA allows studying the relationship between two tables of variables Y and X. While the CCorA is a symmetric method, RDA is non-symmetric. In CCorA, the components extracted from both tables are such that their correlation is maximized. In RDA, the components extracted from X are such that they are as much as possible correlated with the variables of Y. Then, the components of Y are extracted so that they are as much as possible correlated with the components extracted from X.

Principles of RDA

Let Y be a table of response variables with n observations and p variables. This table can be analyzed using Principal Component Analysis (PCA) to obtain a simultaneous map of the observations and the variables in two or three dimensions.

Let X be a table that contains the measures recorded for the same n observations on q quantitative and/or qualitative variables.

Redundancy Analysis allows to obtain a simultaneous representation of the observations, the Y variables, and the X variables in two or three dimensions, that is optimal for a covariance criterion (Ter Braak 1986).

Redundancy Analysis can be divided into two parts:

A constrained analysis in a space which number of dimensions is equal to $min(n-1, p, q)$. This part is the one of main interest as it corresponds to the analysis of the relation between the two tables.

An unconstrained part, which corresponds to the analysis of the residuals. The number of dimensions for the unconstrained RDA is equal to $min(n-1, p)$.

Partial RDA

Partial RDA adds a preliminary step. The X table is subdivided into two groups. The first group $X(1)$ contains conditioning variables which effect we want to remove, as it is either known or without interest for the study. Regressions are run on the Y and $X(2)$ tables and the residuals of the regressions are used for the RDA step. Partial RDA allows to analyze the effect of the second group of variables, after the effect of the first group has been removed.

The terminology Response variables/Observations/Explanatory Variables is used in XLSTAT. When the method is used in ecology, "Species" could be used instead of "Response variables", "Sites" could be used instead of "observations", and "Environmental variables" instead of "Explanatory variables"

Biplot scaling

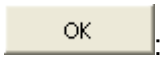
XLSTAT offers three different types of scaling. The type of scaling changes the way the scores of the response variables and the observations are computed, and as a matter of fact, their respective position on the plot. Let $u(ik)$ be the normalized score of variable i on the kth axis, $v(ik)$ the normalized score of observation i on the kth axis, $L(k)$ the eigenvalue corresponding to axis k, and T the total inertia (the sum of the $L(k)$ for the constrained and unconstrained RDA). The three scalings available in XLSTAT are identical to those of vegan (a package for the R software, Oksanen, 2007). The $u(ik)$ are multiplied by c, and the $v(ik)$ by d , and r is a constant equal to $\sqrt[4]{(n-1)T}$, where n is the number of observations.

$$\begin{aligned}
&\text{Scaling 1:} \quad c = r\sqrt{L(k)/T} \qquad d = r \\
&\text{Scaling 2:} \quad c = r \qquad d = r\sqrt{L(k)/T} \\
&\text{Scaling 3:} \quad c = r\sqrt[4]{L(k)/T} \qquad d = r\sqrt[4]{L(k)/T}
\end{aligned}$$

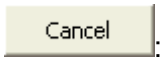
In addition to the observations and the response variables, the explanatory variables can be displayed. The coordinates of the latter are obtained by computing the correlations between the X table and the observation scores.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.


: Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to sites and columns to objects/variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to objects/variables and columns to sites.

General tab:

Response variables Y: Select the table that corresponds to response variables. If the "Column labels" option is activated (column mode) you need to include a header on the first row of the selection. If the "Row labels" option is activated (row mode) you need to include a header in the first column of the selection in the selection.

Explanatory variables X: Select the data that correspond to the various explanatory variables that have been measured for the same observations as for table Y.

- **Quantitative**: Activate this option if you want to use quantitative variables and then select these variables.
- **Qualitative**: Activate this option if you want to use qualitative variables and then select these variables.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Partial RDA: Activate this option to run a partial RDA. If you activate this option, a dialog box will be displayed during the analysis, so that you can select the conditioning variables (see the description for further details).

Column/Row labels: Activate this option if, in column mode, the first row of the selected data contains a header, or in row mode, if the first column of the selected data contains a header.

Observation labels: Activate this option if observation labels are available. Then select the corresponding data. If the "Column labels" option is activated you need to include a header in the selection. If this option is not activated, the sites labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Options tab:

Filter factors: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %**: Activate this option then enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Activate this option to set the number of factors to take into account.

Permutation test: Activate this option if you want to use a permutation test to test the independence of the two tables.

- **Number of permutations**: Enter the number of permutations to perform for the test (Default value: 500)
- **Significance level (%)**: Enter the significance level for the test.

Response variables:

- **Center**: Activate this option to center the variables before running the RDA.
- **Reduce**: Activate this option to standardize the variables before running the RDA

Explanatory variables X:

- **Center**: Activate this option to center the variables before running the RDA.
- **Reduce**: Activate this option to standardize the variables before running the RDA.

Biplot type: Select the type of biplot you want to display. The type changes the way the scores of the response variables and the observations are scaled (see the description for further details).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

RDA results: Activate this option to display the RDA results.

Unconstrained RDA results: Activate this option to display the results of the unconstrained RDA.

Eigenvalues: Activate this option to display the table and the scree plot of the eigenvalues.

Scores (Observations): Activate this option to display the scores of the observations.

Scores (Response variables): Activate this option to display the scores of the response variables.

- **WA scores**: Activate this option to compute and display the weighted average scores.
- **LC scores**: Activate this option to compute and display the linear combinations scores.

Contributions: Activate this option to display the contributions of the observations and the response variables.

Squared cosines: Activate this option to display the squared cosines of the observations and the response variables.

Scores (Explanatory variables): Activate this option to display the scores of the explanatory variables.

Charts tab:

Select the information you want to display on the plot/biplot/triplot.

- **Observations**: Activate this option to display the observations on the chart.
- **Response variables**: Activate this option to display the response variables on the chart.
- **Explanatory variables**: Activate this option to display the explanatory variables on the chart.

Labels: Activate this option to display the labels of the sites on the charts.

- **Colored labels**: Activate this option to display the labels with the same color as the corresponding points. If this option is not activated the labels are displayed in black.

Vectors: Activate this option to display the vectors for the standard coordinates on the asymmetric charts.

- **Length factor**: Activate this option to modulate the length of the vectors.

Results

Summary statistics: This table displays the descriptive statistics for the objects and the explanatory variables.

If a permutation test was requested, its results are first displayed so that we can check if the relationship between the tables is significant or not.

Eigenvalues and percentages of inertia: In these tables are displayed for the constrained RDA and the unconstrained RDA the eigenvalues, the corresponding inertia, and the corresponding percentages, either in terms of constrained inertia (or unconstrained inertia), or in terms of total inertia.

The **scores** of the observations, response variables and explanatory variables are displayed. These coordinates are used to produce the plot.

The chart allows to visualize the relationship between the sites, the objects and the variables. When qualitative variables have been included, the corresponding categories are displayed with a hollowed red circle.

Example

An example of Redundancy Analysis is available on the XLSTAT Help Center:

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

References

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 analysis. An alternative for canonical correlation analysis. *Psychometrika*, **42(2)**, 207-219.

Canonical Correspondence Analysis (CCA)

Use Canonical Correspondence Analysis (CCA) to analyze a contingency table (typically with sites as rows and species in columns) while taking into account the information provided by a set of explanatory variables contained in a second table and measured on the same sites.

In this section:

Description

Canonical Correspondence Analysis (CCA) has been developed to allow ecologists to relate the abundance of species to environmental variables (Ter Braak, 1986). However, this method can be used in other domains. Geomarketing and demographic analyses should be able to take advantage of it.

Principles of CCA

Let T1 be a contingency table corresponding to the counts on n sites of p objects. This table can be analyzed using Correspondence Analysis (CA) to obtain a simultaneous map of the sites and objects in two or three dimensions.

Let T2 be a table that contains the measures recorded on the same n sites of corresponding to q quantitative and/or qualitative variables.

Canonical Correspondence Analysis allows to obtain a simultaneous representation of the sites, the objects, and the variables in two or three dimensions, that is optimal for a variance criterion (Ter Braak 1986, Chessel 1987).

Canonical Correspondence Analysis can be divided into two parts:

A constrained analysis in a space which number of dimensions is equal to q. This part is the one of main interest as it corresponds to the analysis of the relation between the two tables.

An unconstrained part, which corresponds to the analysis of the residuals. The number of dimensions for the unconstrained CCA is equal to min(n-1-q, p-1).

So that a regular CCA can be run, the following constraint must be verified:

Min(n-1-q, p-1)>0.

PLS-CCA allows to avoid this constraint.

Partial CCA

Partial CCA adds a preliminary step. The T2 table is subdivided into two groups of variables: the first group contains conditioning variables which effect we want to remove, as it is either known or without interest for the study. A CCA is ran using these variables. A second CCA is run using the second group of variables which effect we want to analyze. Partial CCA allows to analyze the effect of the second group of variables, after the effect of the first group has been removed.

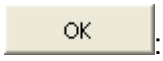
PLS-CCA

Tenenhaus (1998) has shown that it is possible to relate discriminant PLS to CCA. Addinsoft is the first software editor to propose a comprehensive and effective integration between the two methods. Using a restructuring of data based on the proposal Tenenhaus, a PLS step is applied to the data, either to create orthogonal PLS components that are optimally designed for the CCA to avoid the constraints in terms of number of variables that can be used, or to select the most influential variables before running the CCA. As calculations of the CCA step and results are identical to what is done with the classical CCA, users can see this approach as a selection method that identifies the variables of higher interest, either because they are selected in the model, or by looking at the chart of the VIPs (see the section on PLS regression for more information). In the case of a partial CCA, the preliminary step is unchanged.

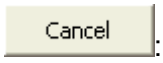
The terminology Sites/Objects/Variables is used in XLSTAT. "Individuals" or "observations" could be used instead of "sites", and "species" instead of "objects" when the method is used in ecology.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down (column mode), XLSTAT considers that rows correspond to sites and columns to objects/variables. If the arrow points to the right (row mode), XLSTAT considers that rows correspond to objects/variables and columns to sites.

General tab:

Sites/Objects data: Select the contingency table that corresponds to the counts of the various objects recorded on each different site. If the "Column labels" option is activated (column mode) you need to include a header on the first row of the selection. If the "Row labels" option is activated (row mode) you need to include a header in the first column of the selection in the selection.

Sites/Variables data: Select the data that correspond to the various explanatory variables that have been measured on the various sites and that you want to use in the analysis.

- **Quantitative**: Activate this option if you want to use quantitative variables and then select these variables.
- **Qualitative**: Activate this option if you want to use qualitative variables and then select these variables.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Partial CCA: Activate this option to run a partial CCA. If you activate this option, a dialog box will be displayed during the analysis, so that you can select the conditioning variables (see the description for additional details).

Column/Row labels: Activate this option if, in column mode, the first row of the selected data contains a header, or in row mode, if the first column of the selected data contains a header.

Sites labels: Activate this option if sites labels are available. Then select the corresponding data. If the "Column labels" option is activated you need to include a header in the selection. If this option is not activated, the sites labels are automatically generated by XLSTAT (Obs1, Obs2 …).

CCA: Activate this option if you want to run a classical CCA.

PLS-CCA: Activate this option if you want to run a PLS-CCA (see the description for additional details).

Options tab:

Filter factors: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %**: Activate this option then enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Activate this option to set the number of factors to take into account.

Permutation test: Activate this option if you want to use a permutation test to test the independence of the two tables.

- **Number of permutations**: Enter the number of permutations to perform for the test (Default value: 500)
- **Significance level (%)**: Enter the significance level for the test.

PLS-CCA: If you choose to run a PLS-CCA the following options are available.

- **Automatic**: Select this option if you want XLSTAT to automatically determine how many PLS components should be used for the CCA step.
- User defined:
    - **Max components**: Activate this option to define the number of components to extract in the PLS step. If this option is not activated, the number of components is automatically determined by XLSTAT.
    - **Number of variables**: Activate this option to define the number of variables that should enter the CCA step. The variables with the higher VIPs are selected. The VIPs that are used are those corresponding to the PLS model with the number of components set in "Max components".

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT does not continue calculations if missing values have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

Row and column profiles: Activate this option to display the row and column profiles.

CCA results: Activate this option to display the CCA results.

Unconstrained CCA results: Activate this option to display the results of the unconstrained CCA.

Eigenvalues: Activate this option to display the table and the scree plot of the eigenvalues.

Principal coordinates: Activate this option to display the principal coordinates of the sites, objects and variables.

Standard coordinates: Activate this option to display the standard coordinates of the sites, objects and variables.

Contributions: Activate this option to display the contributions of the sites, objects and variables.

Squared cosines: Activate this option to display the squared cosines of the sites, objects and variables.

Weighted averages: Activate this option to display the weighted averages that correspond to the variables of the sites/variables table.

Regression coefficients: Activate this option to display regression coefficients that correspond to the various variables in the factor space.

Charts tab:

Sites and objects:

- **Sites and objects / Symmetric**: Activate this option to display a symmetric chart that includes both the sites and the objects. For both the sites and the objects, the principal coordinates of are used.

- **Sites / Asymmetric**: Activate this option to display the asymmetric chart of the sites. The principal coordinates are used for the sites, and the standard coordinates are used for the objects.
- **Objects / Asymmetric**: Activate this option to display the asymmetric chart of the objects. The principal coordinates are used for the objects, and the standard coordinates are used for the sites.
- **Sites**: Activate this option to display a chart on which only the sites are displayed. The principal coordinates are used.
- **Objects**: Activate this option to display a chart on which only the objects are displayed. The principal coordinates are used.

Variables:

- **Correlations**: Activate this option to display the quantitative and qualitative variables on the charts, using as coordinates their correlations (equal to their standard coordinates).
- **Regression coefficients**: Activate this option to display the quantitative and qualitative variables on the charts, using the regression coefficients as coordinates.

Labels: Activate this option to display the labels of the sites on the charts.

- **Colored labels**: Activate this option to display the labels with the same color as the corresponding points. If this option is not activated the labels are displayed in black.

Vectors: Activate this option to display the vectors for the standard coordinates on the asymmetric charts.

- **Length factor**: Activate this option to modulate the length of the vectors.

Results

Summary statistics: This table displays the descriptive statistics for the objects and the explanatory variables.

Inertia: This table displays the distribution of the inertia between the constrained CCA and the unconstrained CCA.

Eigenvalues and percentages of inertia: In these tables are displayed for the CCA and the unconstrained CCA the eigenvalues, the corresponding inertia, and the corresponding percentages, either in terms of constrained inertia (or unconstrained inertia), or in terms of total inertia.

Weighted averages: This table displays the weighted means as well the global weighted means.

The **principal coordinates** and **standard coordinates** of the sites, the objects and the variables are then displayed. These coordinates are used to produce the various charts.

Regression coefficients: This table displays the regression coefficients of the variables in the factor space.

The charts allow to visualize the relationship between the sites, the objects and the variables. When qualitative variables have been included, the corresponding categories are displayed with a hollowed red circle.

Example

An example of Canonical Correspondence Analysis is available on XLSTAT Help Center:

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

References

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.

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.

Principal Coordinate Analysis (PCoA)

Use Principal Coordinate Analysis to graphically visualize a square matrix that describes the similarity or the dissimilarity between p elements (individuals, variables, objects, …).

In this section:

Description

Principal Coordinate Analysis (often referred to as PCoA) is aimed at graphically representing a resemblance matrix between p elements (individuals, variables, objects, …).

If the input matrix is a similarity matrix, XLSTAT transforms it into a dissimilarity matrix before applying the calculations described by Gower (1966), or before any of the changes suggested by various authors and summarized in the *Numerical Ecology* book by Legendre and Legendre (1998).

Concept

Let D be a $p \times p$ symmetric matrix that contains the distances between p elements: we compute the A matrix which elements $a(ij)$, corresponding to the ith row and to the jth column, are given by

$$a(ij) = -d^2(ij)/2$$

We then center the A matrix by rows and by columns to obtain the Δ_1 matrix which elements $\delta_1(ij)$ are given by

$$\delta_1(ij) = a(ij) - \overline{a}(i) - \overline{a}(j) - \overline{a}$$

where $\overline{a}(i)$ is the mean of the $a(ij)$ for row i, $\overline{a}(j)$ is the mean of the $a(ij)$ for column j and $\overline{a}$ is the mean of all the elements.

Last, we compute the eigen-decomposition of Δ_1. The eigenvectors are sorted by decreasing order of eigenvalues and transformed so that, if $u(k)$ is the eigenvector associated to the $\lambda(k)$ eigenvalue, we have:

$$u'(k)u(k) = \lambda(k)$$

The rescaled eigenvectors correspond to the principal coordinates that can be used to display the p objects in a space with 1, 2, $p-1$ dimensions.

As with PCA (Principal Component Analysis) eigenvalues can be interpreted in terms of percentage of total variability that is being represented in a reduced space.

Note: because Δ_1 is centered, we obtain at most, $p-1$ non null eigenvalues. In the case where the initial matrix D is an Euclidean matrix distance, we can easily understand that $p-1$ axes are enough to fully describe p objects (by two points passes one line, three points are contained in a plane, …). In the case where the points are confounded in a sub-space, we can obtain several null eigenvalues (for example, three points can be aligned on a line).

Negative eigenvalues

When the D matrix is not metric, or if missing values were present in the data that were used to compute the distances, the eigen-decomposition can lead to negative eigenvalues. This can especially happen with semi-metric or non metric distances. This problem is described in the article by Gower and Legendre (1986).

XLSTAT suggests two transformations to solve the problem of negative eigenvalues. The first solution consists in taking as input distances the square root of the input distances. The second, inspired by the results of Lingoes (1971), consists in adding a constant to the D matrix (except to the diagonal elements) such that there is no negative eigenvalue. This constant is equal to the opposite of the largest negative eigenvalue.

PCA, MDS and PCoA

PCA and PCoA are quite similar in that PCA can also represent observations in a space with less dimensions, the later being optimal in terms of variability carried. A PCoA applied to matrix of Euclidean distances between observations (calculated after standardization of the columns using the unbiased standard deviation) leads to the same results as a PCA based on the correlation matrix. The eigenvalues obtained with the PCoA are equal to $(p-1)$ times those obtained with the PCA.

PCoA and MDS (Multidimensional Scaling) share the same goal of representing objects for which we have a proximity matrix.

MDS has two drawbacks compared with PCoA:

- The algorithm is much more complex and performs slower.
- Axes obtained with MDS cannot be interpreted in terms of variability.

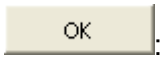
MDS has two advantages compared with PCoA:

- The algorithm allows having missing data in the proximity matrix.

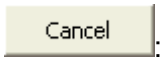
- The non-metric version of MDS provides a simpler and clear way to handle matrices where only the ranking of the distances is important.

Dialog box

The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.

: Click this button to display the help.


: Click this button to reload the default options.


: Click this button to delete the data selections.

General tab:

Data: Select a similarity or dissimilarity matrix. If only the lower or upper triangle is available, the table is accepted. If differences are detected between the lower and upper parts of the selected matrix, XLSTAT warns you and offers to change the data (by calculating the average of the two parts) to continue with the calculations.

Dissimilarities / Similarities: Choose the option that corresponds to the type of your data.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet in the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Labels included: Activate this option if you have included row and column labels in the selection.

Options tab:

Correction for negative eigenvalues: Activate the option that corresponds to the strategy to apply if eigenvalues are detected during the eigen- decomposition:

- **None**: Nothing is done when negative eigenvalues are found.
- **Square root**: The elements of the distance matrix D are replaced by their square root.
- **Lingoes**: A transformation is applied so that that eigen-decomposition does not lead to any negative eigenvalue.

Filter factors: You can activate one of the following two options in order to reduce the number of factors for which results are displayed.

- **Minimum %**: Activate this option then enter the minimum percentage of the total variability that the chosen factors must represent.
- **Maximum Number**: Activate this option to set the number of factors to take into account.

Outputs tab:

Delta1 matrix: Activate this option to display the Δ_1 matrix that is used to compute the eigenvalues and the eigenvectors.

Eigenvalues: Activate this option to display the table and the chart (scree plot) of the eigenvalues.

Principal coordinates: Activate this option to display the principal coordinates.

Contributions: Activate this option to display the contributions.

Squared cosines: Activate this option to display the squared cosines.

Charts tab:

Chart: Activate this option to display the chart.

Results

Delta1 matrix: This matrix corresponds to the Δ_1 matrix of Gower, used to compute the eigen-decomposition.

Eigenvalues and percentage of inertia: this table displays the eigenvalues and the corresponding percentage of inertia.

Principal coordinates: This table displays of the principal coordinates of the objects that are used to create the chart where the proximities between the charts can be interpreted.

Contributions: This table displays the contributions that help evaluate how much an object contributes to a given axis.

Squared cosines: This table displays the contributions that help evaluate how close an object is to a given axis.

Example

An example showing how to run a Principal Coordinate Analysis is available on the XLSTAT Help Center at:

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

References

Cailliez F. and Pagè s 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 is a module of XLSTAT that is dedicated to the component based structural equation modeling in particular with methods such as Partial Least Squares Path Modeling (PLS-PM / PLS-SEM), Generalized Structured Components Analysis (GSCA) or Regularized Generalized Canonical Correlation Analysis (RGCCA). These are innovative methods for representing complex relationships between observed variables and latent variables.

The XLSTAT-PLSPM methods can be used in different fields such as in marketing for the analysis of consumer satisfaction. 3 levels of display are proposed (Classic, Expert, Marketing) in order to adapt to different types of users.

In this section:

Description

Partial Least Squares Path Modeling (PLS-PM) is a statistical approach for modeling complex multivariable relationships (structural equation models) among observed and latent variables. Since a few years, this approach has been enjoying increasing popularity in several sciences (Esposito Vinzi et al., 2007). Structural Equation Models include a number of statistical methodologies allowing the estimation of a causal theoretical network of relationships linking latent complex concepts, each measured by means of a number of observable indicators.

The first presentation of the finalized PLS approach to path models with latent variables has been published by Wold in 1979 and then the main references on the PLS algorithm are Wold (1982 and 1985).

Herman Wold opposed LISREL (Jöreskog, 1970) "hard modeling" (heavy distribution assumptions, several hundreds of cases necessary) to PLS "soft modeling" (very few distribution assumptions, few cases can suffice). These two approaches to Structural Equation Modeling have been compared in Jöreskog and Wold (1982).

From the standpoint of structural equation modeling, PLS-PM is a component- based approach where the concept of causality is formulated in terms of linear conditional expectation. PLS-PM seeks for optimal linear predictive relationships rather than for causal mechanisms thus privileging a prediction- relevance oriented discovery process to the statistical testing of causal hypotheses. Two very important review papers on PLS approach to Structural Equation Modeling are Chin (1998, more application oriented) and Tenenhaus et al. (2005, more theory oriented).

Furthermore, PLS Path Modeling can be used for analyzing multiple tables and it is directly related to more classical data analysis methods used in this field. In fact, PLS-PM may be also viewed as a very flexible approach to multi-block (or multiple table) analysis by means of both the hierarchical PLS path model and the confirmatory PLS path model (Tenenhaus and Hanafi, 2007). This approach clearly shows how the "data-driven" tradition of multiple table analysis can be somehow merged in the "theory-driven" tradition of structural equation modeling so as to allow running the analysis of multi-block data in light of current knowledge on conceptual relationships between tables.

Other methods such as Generalized Structured Components Analysis (GSCA) or Regularized Generalized Canonical Correlation Analysis (RGCCA) have been introduced to tackle the weakness of PLS-PM.

The PLS Path Modeling algorithm

A PLS Path model is described by two models: (1) a measurement model relating the manifest variables to their own latent variable and (2) a structural model relating some endogenous latent variables to other latent variables. The measurement model is also called the outer model and the structural model the inner model.

1. Manifest variables standardization

There exist four options for the standardization of the manifest variables depending upon three conditions that eventually hold in the data:

- Condition 1: The scales of the manifest variables are comparable. For instance, in the ECSI example the item values (between 0 and 100) are comparable. On the other hand, for instance, *weight* in tons and *speed* in km/h would not be comparable.
- Condition 2: The means of the manifest variables are interpretable. For instance, if the difference between two manifest variables is not interpretable, the location parameters are meaningless.
- Condition 3: The variances of the manifest variables reflect their importance.

If condition 1 does not hold, then the manifest variables have to be standardized (mean 0 and variance 1).

If condition 1 holds, it is useful to get the results based on the raw data. But the calculation of the model parameters depends upon the validity of the other conditions:

Condition 2 and 3 do not hold: The manifest variables are standardized (mean 0 variance 1) for the parameter estimation phase. Then the manifest variables are rescaled to their original means and variances for the final expression of the weights and loadings.

Condition 2 holds, but not condition 3: The manifest variables are not centered, but are standardized to unitary variance for the parameter estimation phase. Then the manifest variables are rescaled to their original variances for the final expression of the weights and loadings (to be defined later).

Conditions 2 and 3 hold: Use the original variables.

Lohmöller (1989) introduced a standardization parameter to select one of these four options:

Variable scales are comparable	Means are interpretable	Variance is related to variable importance	Mean	Variance	Rescaling	METRIC
No			0	1	No	1
Yes	No	No	0	1	Yes	2
Yes	Yes	No	Original	1	Yes	3
Yes	Yes	Yes	Original	Original		4

With METRIC=1 being "Standardized, weights on standardized MV", METRIC=2 being "Standardized, weights on raw MV", METRIC=3 being "Reduced, weights on raw MV" and METRIC=4 being "Raw MV".

2. The measurement model

A latent variable (LV) ξ is an unobservable variable (or construct) indirectly described by a block of observable variables x_h which are called manifest variables (MV) or indicators. There are three ways to relate the manifest variables to their latent variables, respectively called the reflective way, the formative one, and the MIMIC (Multiple effect Indicators for MultIple Causes) way.

2.1. The reflective way

2.1.1. Definition

In this model each manifest variable reflects its latent variable. Each manifest variable is related to its latent variable by a simple regression:

$$x_h = \pi_{h0} + \pi_h \xi + \epsilon_h \quad (1)$$

where ξ has mean m and standard deviation 1. It is a reflective scheme: each manifest variable x_h reflects its latent variable ξ. The only hypothesis made on model (1) is called by H. Wold the *predictor specification* condition:

$$E(x_h|\xi) = \pi_{h_0} + \pi_h \xi \quad (2)$$

This hypothesis implied that the residual ϵ_h has a zero mean and is uncorrelated with the latent variable ξ.

2.1.2. Check for unidimensionality

In the reflective way the block of manifest variables is unidimensional in the meaning of factor analysis. On practical data this condition has to be checked. Three main tools are available to check the unidimensionality of a block: use of principal component analysis of each block of manifest variables, Cronbach's α and Dillon-Goldstein's ρ.

a) Principal component analysis of a block

A block is essentially unidimensional if the first eigenvalue of the correlation matrix of the block MVs is larger than 1 and the second one smaller than 1, or at least very far from the first one. The first principal component can be built in such a way that it is positively correlated with all (or at least a majority of) the MVs. There is a problem with MV negatively correlated with the first principal component.

b) Cronbach's α

Cronbach's a can be used to check unidimensionality of a block of p variables x_h when they are all positively correlated. Cronbach has proposed the following procedure for *standardized* variables:

$$\alpha = \frac{\sum_{h \neq h'} cor(x_h, x_{h'})}{p + \sum_{h \neq h'} cor(x_h, x_{h'})} \times \frac{p}{p-1} \tag{3}$$

The Cronbach's α is also defined for original (raw) variables as:

$$\alpha = \frac{\sum_{h \neq h'} cov(x_h, x_{h'})}{var\left(\sum_h x_h\right)} \times \frac{p}{p-1} \tag{4}$$

A block is considered as unidimensional when the Cronbach's α is larger than 0.7.

c) Dillon-Goldstein's ρ

The sign of the correlation between each MV x_h and its LV ξ is known by construction of the item and is supposed here to be positive. In equation (1) this hypothesis means that all the loadings π_h are positive. A block is unidimensional if all these loadings are large.

The Goldstein-Dillon's ρ is defined by:

$$\rho = \frac{\left(\sum_{h=1}^{p} \pi_h\right)^2 \mathrm{Var}(\xi)}{\left(\sum_{h=1}^{p} \pi_h\right)^2 \mathrm{Var}(\xi) + \sum_{h=1}^{p} Var\left(\epsilon_h\right)} \tag{5}$$

Let's now suppose that all the MVs x_h and the latent variable ξ are standardized. An approximation of the latent variable ξ is obtained by standardization of the first principal component t_1 of the block MVs. Then π_h is estimated by $cor(x_h, t_1)$ and, using equation (1), $Var(\epsilon_h)$ is estimated by $1 - cor^2(x_h, t_1)$. So we get an estimate of the Dillon-Goldstein's ρ:

$$\hat{\rho} = \frac{\left[\sum_{h=1}^{p} \mathrm{cor}\left(x_h, t_1\right)\right]^2}{\left[\sum_{h=1}^{p} \mathrm{cor}\left(x_h, t_1\right)\right]^2 + \sum_{h=1}^{p} \left[1 - \mathrm{cor}^2\left(x_h, t_1\right)\right]} \tag{6}$$

A block is considered as unidimensional when the Dillon-Goldstein's $\hat{\rho}$ is larger than 0.7. This statistic is considered to be a better indicator of the unidimensionality of a block than the Cronbach's α (Chin, 1998, p.320).

PLS Path Modeling is a mixture of *a priori* knowledge and data analysis. In the reflective way, the a priori knowledge concerns the unidimensionality of the block and the signs of the loadings. The data have to fit this model. If they do not, they can be modified by removing some manifest variables that are far from the model. Another solution is to change the model and use the formative way that will now be described.

2.2. The formative way

In the formative way, it is supposed that the latent variable ξ is generated by its own manifest variables. The latent variable ξ is a linear function of its manifest variables plus a residual term:

$$\xi = \sum_h \varpi_h x_h + \delta \tag{7}$$

In the formative model the block of manifest variables can be multidimensional. The predictor specification condition is supposed to hold as:

$$E\left(\xi | x_1, \ldots, x_{p_j}\right) = \sum_h \varpi_h x_h \tag{8}$$

This hypothesis implies that the residual vector δ has a zero mean and is uncorrelated with the MVs x_h.

2.3. The MIMIC way

The MIMIC way is a mixture of the reflective and formative ways.

The measurement model for a block is the following:

$$x_h = \pi_{h_0} + \pi_h \xi + \epsilon_h \text{ pour } h = 1, \ldots, p_1 \tag{9}$$

where the latent variable is defined by:

$$\xi = \sum_{h=p_1+1}^{p} \varpi_h x_h + \delta \tag{10}$$

The p_1 first manifest variables follow a reflective way and the $\left(p - p_1\right)$ last ones a formative way. The predictor specification hypotheses still hold and lead to the same consequences as before on the residuals.

3. The structural model

The causality model leads to linear equations relating the latent variables between them (the structural or inner model):

$$\xi_j = \beta_{j_0} + \sum_i \beta_{ji}\xi_i + v_j \tag{11}$$

The predictor specification hypothesis is still applied.

A latent variable, which never appears as a dependent variable, is called an exogenous variable. Otherwise it is called an endogenous variable.

4. The Estimation Algorithm

4.1. Latent variables Estimation

The latent variables ξ are estimated according to the following procedure.

4.1.1. Outer estimate y_j of the standardized latent variable ($\xi_j - m_j$)

The standardized latent variables (mean = 0 and standard deviation = 1) are estimated as linear combinations of their centered manifest variables:

$$y_j \propto \pm \left[\sum w_{jh}\left(x_{jh} - \overline{x}_{jh}\right)\right] \tag{12}$$

where the symbol "$\propto$" means that the left variable represents the standardized right variable and the "$\pm$" sign shows the sign ambiguity. This ambiguity is solved by choosing the sign making y_j positively correlated to a majority of x_{jh}.

The standardized latent variable is finally written as:

$$y_j = \sum \tilde{w}_{jh}(x_{jh} - \overline{x}_{jh}) \tag{13}$$

The coefficients w_{jh} and $\tilde{w_{jh}}$ are both called the outer weights.

The mean m_j is estimated by:

$$\hat{m}_j = \sum \tilde{w}_{jh}\overline{x}_{jh} \tag{14}$$

and the latent variable ξ_j by:

$$\hat{\xi}_j = \sum \tilde{w}_{jh}x_{jh} = y_j + \hat{m}_j \tag{15}$$

When all manifest variables are observed on the same measurement scale, it is nice to express (Fornell (1992)) latent variables estimates in the original scale as:

$$\hat{\xi}_j^{\star} = \frac{\sum \tilde{w}_{jh}x_{jh}}{\sum \tilde{w}_{jh}} \tag{16}$$

Equation (16) is feasible when all outer weights are positive. Finally, most often in real applications, latent variables estimates are required on a 0-100 scale so as to have a reference scale to compare individual scores. From the equation (16), for the i-th observed case, this is easily obtained by the following transformation:

$$\hat{\xi}_j^{0-100} = 100 \times \frac{\hat{\xi}_j^{\star} - x_{min}}{x_{max} - x_{min}} \tag{17}$$

where x_{min} and x_{max} are, respectively, the minimum and the maximum value of the measurement scale common to all manifest variables.

4.1.2. Inner estimate z_j of the standardized latent variable $(x_j - m_j)$

The inner estimate z_j of the standardized latent variable $(x_j - m_j)$ is defined by:

$$z_j \propto \sum_{j':\xi_{j'} \text{ is connected to } \xi_j} e_{jj'} y_{j'} \tag{18}$$

where the inner weights $e_{jj'}$ are equal to the signs of the correlations between y_j and the y'_j connected with y_j. Two latent variables are connected if there exists a link between the two variables: an arrow goes from one variable to the other in the arrow diagram describing the causality model. This choice of inner weights is called the *centroid scheme*.

Centroid scheme :

This choice shows a drawback in case the correlation is approximately zero as its sign may change for very small fluctuations. But it does not seem to be a problem in practical applications.

In the original algorithm, the inner estimate is the right term of (18) and there is no standardization. We prefer to standardize because it does not change anything for the final inner estimate of the latent variables and it simplifies the writing of some equations.

Two other schemes for choosing the inner weights exist: the factorial scheme and the path weighting (or structural) scheme. These two new schemes are defined as follows:

Factorial scheme :

The inner weights e_{ji} are equal to the correlation between y_i and y_j. This is an answer to the drawbacks of the centroid scheme described above.

Path weighting scheme (structural) :

The latent variables connected to ξ_j are divided into two groups: the predecessors of ξ_j, which are latent variables explaining ξ_j, and the followers, which are latent variables explained by ξ_j.

For a predecessor ξ'_j of the latent variable ξ_j, the inner weight $e_{jj'}$ is equal to the regression coefficient of y'_j in the multiple regression of y_j on all the y'_j related to the predecessors of ξ_j. If ξ'_j is a successor of ξ_j then the inner weight $e_{jj'}$ is equal to the correlation between y'_j and y_j.

These new schemes do not significantly influence the results but are very important for theoretical reasons. In fact, they allow to relate PLS Path modeling to usual multiple table analysis methods.

The Horst scheme :

The internal weights e_{ji} are always 1. This is one of the first scheme developed for the PLS Path Modeling.

4.2. The PLS algorithm for estimating the weights

4.2.1. Estimation modes for the weights w_{jh}

There are three classical ways to estimate the weights w_{jh}: Mode A, Mode B and Mode C.

Mode A:

In mode A the weight w_{jh} is the regression coefficient of z_j in the simple regression of x_{jh} on the inner estimate z_j:

$$w_{jh} = cov(x_{jh}, z_{jh}) \tag{19}$$

as z_j is standardized.

Mode B:

In mode B the vector w_j of weights w_{jh} is the regression coefficient vector in the multiple regression of z_j on the manifest centered variables ($x_{jh} - \overline{x}_{jh}$) related to the same latent variable ξ_j:

$$w_{jh} = \left({X_j}' - X_j\right)^{-1} {X_j}' z_j \tag{20}$$

where X_j is the matrix with columns defined by the centered manifest variables $x_{jh} - \overline{x}_{jh}$ related to the j-th latent variable ξ_j.

Mode A is appropriate for a block with a reflective measurement model and Mode B for a formative one. Mode A is often used for an endogenous latent variable and mode B for an exogenous one. Modes A and B can be used simultaneously when the measurement model is the MIMIC one. Mode A is used for the reflective part of the model and Mode B for the formative part.

In practical situations, mode B is not so easy to use because there is often strong multicollinearity inside each block. When this is the case, PLS regression may be used instead of OLS multiple regression. As a matter of fact, it may be noticed that mode A consists in taking

the first component from a PLS regression, while mode B takes all PLS regression components (and thus coincides with OLS multiple regression). Therefore, running a PLS regression and retaining a certain number of significant components may be meant as a new intermediate mode between mode A and mode B.

Mode C (centroid):

In mode C the weights are all equal in absolute value and reflect the signs of the correlations between the manifest variables and their latent variables:

$$w_{jh} = sign(cor(x_{jh}, z_j)) \tag{21}$$

These weights are then normalized so that the resulting latent variable has unitary variance. Mode C actually refers to a formative way of linking manifest variables to their latent variables and represents a specific case of Mode B whose comprehension is very intuitive to practitioners.

4.2.2. Estimating the weights

The starting step of the PLS algorithm consists in beginning with an arbitrary vector of weights w_{jh}. These weights are then standardized in order to obtain latent variables with unitary variance.

A good choice for the initial weight values is to take $w_{jh} = sign(cor(x_{jh}, \xi_h))$ or, more simply, $w_{jh} = sign(cor(x_{jh}, \xi_h))$ for $h = 1$ and 0 otherwise or they might be the elements of the first eigenvector from a PCA of each block.

Then the steps for the outer and the inner estimates, depending on the selected mode, are iterated until convergence (guaranteed only for the two- blocks case, but practically always encountered in practice even with more than two blocks).

After the last step, final results are yielded for the inner weights $\tilde{w_{jh}}$, the standardized latent variable $y_j = \sum \tilde{w}_{jh}\left(x_{jh} - \overline{x}_{jh}\right)$, the estimated mean $\hat{m}_j = \sum \tilde{w}_{jh}\overline{x}_{jh}$ of the latent variable ξ_j, and the final estimate $\hat{\xi}_j = \sum \tilde{w}_{jh}x_{jh} = y_j + \hat{m}_j$ of ξ_j. The latter estimate can be rescaled according to transformations (16) and (17).

The latent variable estimates are sensitive to the scaling of the manifest variables in Mode A, but not in mode B. In the latter case, the outer LV estimate is the projection of the inner LV estimate on the *space* generated by its manifest variables.

4.3. Estimation of the structural equations

The structural equations (11) are estimated by individual OLS multiple regressions where the latent variables ξ_j are replaced by their estimates $\hat{\xi}_j$. As usual, the use of OLS multiple regressions may be disturbed by the presence of strong multicollinearity between the estimated latent variables. In such a case, PLS regression may be applied instead.

5. Missing Data Treatment

In XLSTAT-PLSPM, there exists a specific treatment for missing data (Lohmöller, 1989):

1. When some cells are missing in the data, means and standard deviations of the manifest variables are computed on all the available data.

2. All the manifest variables are centered.

3. If a unit has missing values on a whole block j, the value of the latent variable estimate y_j is missing for this unit.

4. If a unit i has some missing values on a block j (but not all), then the outer estimate y_{ji} is defined by:

$$y_{ji} = \sum_{jh:x_{jhi} \text{ exists}} \tilde{w}_{jh}(x_{jhi} - \overline{x}_{jh})$$

That means that each missing data of variable x_{jh} is replaced by the mean $\overline{x}_{jh}$.

5. If a unit i has some missing values on its latent variables, then the inner estimate z_{ji} is defined by:

$$z_{ji} = \sum_{k:\xi_k \text{ is connected with } \xi_j \text{ and } y_{ki} \text{ exists}} e_{jk} y_{ki}$$

That means that each missing data of variable y_k is replaced by its mean 0.

6. The weights w_{jh} are computed using all the available data on the basis of the following procedures:

- For mode A: The outer weight w_{jh} is the regression coefficient of z_j in the regression of $(x_{jh} - \overline{x}_{jh})$ on z_j calculated on the available data.
- For mode B: When there are no missing data, the outer weight vector w_j is equal to:

$$w_j = [Var(X_j)]^{-1} Cov(X_j, z_j)$$

where $Var(X_j)$ is the covariance matrix of X_j and $Cov(X_j, z_j)$ the column vector of the covariances between the variables x_{jh} and z_j.

When there are missing data, each element of $Var(X_j)$ and $Cov(X_j, z_j)$ is computed using all the pairwise available data and w_j is computed using the previous formula.

This pairwise deletion procedure shows the drawback of possibly computing covariances on different sample sizes and/or different statistical units. However, in the case of few missing values, it seems to be very robust. This justifies why the blindfolding procedure that will be presented in the next section yields very small standard deviations for parameters.

7. The path coefficients are the regression coefficients in the multiple regressions relating some latent variables to some others. When there are some missing values, the procedure described in point 6 (Mode B) is also used to estimate path coefficients.

Nevertheless, missing data can be also treated with other classical procedures, such as mean imputation, listwise deletion, multiple imputation, the NIPALS algorithm (discussed below) and so on so forth.

6. Model Validation

A path model can be validated at three levels: (1) the quality of the measurement model, (2) the quality of the structural model, and (3) each structural regression equation.

6.1. Communality and redundancy

The communality index measures the quality of the measurement model for each block. It is defined, for block j, as:

$$Communality_j = \frac{1}{p_j}\sum_{h=1}^{p_j} cor^2(x_{jh}, y_j) \tag{22}$$

The average communality is the average of all the $\mathrm{cor}^2(x_{jh}, y_j)$:

$$\overline{Communality} = \frac{1}{p}\sum_{j=1}^{J} p_j Communality_j \tag{23}$$

where p is total number of manifest variables in all blocks.

The redundancy index measures the quality of the structural model for each endogenous block. It is defined, for an endogenous block j, as:

$$Redundancy_j = Communality_j \times R^2(y_j, \{y_{j'} \text{ explaining } y_j\}) \tag{24}$$

The average redundancy for all endogenous blocks can also be computed.

A global criterion of goodness-of-fit (GoF) can be proposed (Amato, Esposito Vinzi and Tenenhaus 2004) as the geometric mean of the average communality and the average R^2:

$$GoF = \sqrt{\overline{\text{Communality}} \times \overline{R^2}} \tag{25}$$

As a matter of fact, differently from LISREL, PLS Path Modeling does not optimize any global scalar function so that it naturally lacks of an index that can provide the user with a global validation of the model (as it is instead the case with c^2 and related measures in LISREL). The GoF represents an operational solution to this problem as it may be meant as an index for validating the PLS model globally, as looking for a compromise between the performances of the measurement and the structural model, respectively.

6.2. The Blindfolding approach: cross-validated communality and redundancy

The cv-communality (cv stands for cross-validated) index measures the quality of the measurement model for each block. It is a kind of cross-validated R-square between the block MVs and their own latent variable calculated by a blindfolding procedure.

The quality of each structural equation is measured by the cv-redundancy index (i.e. Stone-Geisser's Q^2). It is a kind of cross-validated R-square between the manifest variables of an endogenous latent variable and all the manifest variables associated with the latent variables explaining the endogenous latent variable, using the estimated structural model.

Following Wold (1982, p. 30), the cross-validation test of Stone and Geisser fits soft modeling like hand in glove. In PLS Path Modeling statistics on each block and on each structural regression are available.

The significance levels of the regression coefficients can be computed using the usual Student's t statistic or using cross-validation methods like jack- knife or bootstrap.

Here is the description of the blindfolding approach proposed by Herman Wold:.

1. The data matrix is divided into G groups. The value $G = 7$ is recommended by Herman Wold. We give in the following table an example on a dataset made by 12 statistical units and 5 variables. The first group is related to letter a, the second one to letter b, and so on.

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. Each group of cells is removed at its turn from the data. So a group of cells appears to be missing (for example all cells with letter a).

3. A PLS model is run G times by excluding each time one of the groups.

4. One way to evaluate the quality of the model consists in measuring its capacity to predict manifest variables using other latent variables. Two indices are used: communality and redundancy.

5. In the communality option, we get prediction for the values of the centered manifest variables not included in the analysis, using the latent variable estimate, by the following formula:

$$Pred(x_{jhi} - \overline{x}_{jh}) = \hat{\pi}_{jh(-i)} y_{j(-i)}$$

where $\hat{\pi}_{jh(-i)}$ and $y_{j(-i)}$ are computed on data where the i-th value of variable x_{jh} is missing.

The following terms are computed:

- Sum of squares of observations for one MV: $SSO_{jh} = \sum_i (x_{jhi} - \overline{x}_{jh})^2$.
- Sum of squared prediction errors for one MV: $SSE_{jh} = \sum_i (x_{jhi} - \overline{x}_{jh} - \hat{\pi}_{jh(-i)} y_{j(-i)})^2$.
- Sum of squares of observations for Block j: $SSO_j = \sum_h SSO_{jh}$.

- Sum of squared prediction errors for Block j: $SSE_j = \sum_h SSE_{jh}$.
- CV-Communality measure for Block j: $H_j^2 = 1 - \frac{SSE_j}{SSO_i}$.

6. The index H_j^2 is the cross-validated communality index. The mean of the cv-communality indices can be used to measure the global quality of the measurement model if they are positive for all blocks.

In the redundancy option, we get a prediction for the values of the centred manifest variables not used in the analysis by using the following formula:

$$Pred(x_{jhi} - \overline{x}_{jh}) = \hat{\pi}_{jh(-i)} Pred(y_{j(-i)})$$

where $\hat{\pi}_{jh(-i)}$ is the same as in the previous paragraph and $Pred(y_{j(-i)})$ is the prediction for the i-th observation of the *endogenous* latent variable y_j using the regression model computed on data where the i-th value of variable x_{jh} is missing.

The following terms are also computed:

- Sum of squared prediction errors for one MV:

$$SSE'_{jh} = \sum_i (x_{jhi} - \overline{x}_{jh} - \hat{\pi}_{jh(-i)} Pred(y_{j(-i)}))^2$$

- Sum of squared prediction errors for block j:

$$SSE'_j = \sum_h SSE'_{jh}$$

- CV-Redundancy measure for an endogenous block j:

$$F_j^2 = 1 - \frac{SSE'_j}{SSO_j}$$

The index F_j^2 is the cross-validated redundancy index. The mean of the various cv-redundancy indices related to the endogenous blocks can be used to measure the global quality of the structural model if they are positive for all endogenous blocks.

6.3. Resampling: Jackknife and Bootstrap

The significance of PLS-PM parameters, coherently with the distribution-free nature of the estimation method, is assessed by means of non-parametric procedures. As a matter of fact, besides the classical blindfolding procedure, Jackknife and Bootstrap resampling options are available.

6.3.1. Jackknife

The Jackknife procedure builds resamples by deleting a certain number of units from the original sample (with size N). The default option consists in deleting 1 unit at a time so that each Jackknife sub-sample is made of $N-1$ units. Increasing the number of deleted units leads to a potential loss in robustness of the t-statistic because of a smaller number of sub-samples. The complete statistical procedure is described in Chin (1998, p.318-320).

6.3.2. Bootstrap

The Bootstrap samples, instead, are built by resampling with replacement from the original sample. The procedure produces samples consisting of the same number of units as in the original sample. The number of resamples has to be specified. The default is 100 but a higher number (such as 200) may lead to more reasonable standard error estimates.

We must take into account that, in PLS-PM, latent variables are defined up to the sign. It means that $y_j = \sum \tilde{w}_{jh}(x_{jh} - \overline{x}_{jh})$ and $-y_j$ are both equivalent solutions. In order to remove this indeterminacy, Wold (1985) suggests retaining the solution where the correlations between the manifest variables x_{jh} and the latent variable y_j show a majority of positive signs. Referring to the signs of the elements in the first eigenvector obtained on the original sample is also a way of controlling the sign in the different bootstrap re-samples.

GSCA (Generalized Structured Component Analysis)

This method introduced by Hwang and Takane (2011), allows to optimize a global function using an algorithm called Alternating Least Square algorithm (ALS).

GSCA lies in the tradition of component analysis. It substitutes components for factors as in PLS. Unlike PLS, however, GSCA offers a global least squares optimization criterion, which is consistently minimized to obtain the estimates of model parameters. GSCA is thus equipped with an overall measure of model fit while fully maintaining all the advantages of PLS (e.g., less restricted distributional assumptions, no improper solutions, and unique component score estimates). In addition, GSCA handles more diverse path analyses, compared to PLS.

Let Z denote an N by J matrix of observed variables. Assume that Z is columnwise centered and scaled to unit variance. Then, the model for GSCA may be expressed as

$$ZV = ZWA + E$$

and

$$P = GA + E \tag{1}$$

where $P = ZV$, and $G = ZW$. In (1), P is an N by T matrix of all endogenous observed and composite variables, G is an N by D matrix of all exogenous observed and composite variables, V is a J by T matrix of component weights associated with the endogenous variables, W is a J by D matrix of component weights for the exogenous variables, A is a D by T supermatrix consisting of a matrix of component loadings relating components to their

observed variables, denoted by C, in addition to a matrix of path coefficients between components, denoted by B, that is, $A = [C, B]$, and E is a matrix of residuals.

We estimate the unknown parameters V,W, and A in such a way that the sum of squares of the residuals, $E = ZV - ZWA = P - GA$, is as small as possible. This amounts to minimizing

$$f = SS(ZV - ZWA) = SS(P - GA) \quad (2)$$

with respect to V, W, and A, where $SS(X) = trace(X'X)$. The components in P and/or G are subject to normalization for identification purposes.

We cannot solve (2) in an analytic way since V, W, and A can comprise zero or any fixed elements. Instead, we develop an alternating least squares (ALS) algorithm (de Leeuw, Young, & Takane, 1976) to minimize (2). In general, ALS can be viewed as a special type of the FP algorithm where the fixed point is a stationary (accumulation) point of a function to be optimized.

The proposed ALS algorithm consists of two steps: In the first step, A is updated for fixed V and W. In the second step, V and W are updated for fixed A. (Hwang and Takane, 2004)

RGCCA (Regularized Generalized Canonical Correlation Analysis)

This method introduced by Tenenhaus et al. (2011), allows to optimize a global function using an algorithm very similar to the PLSPM algorithm.

Unlike the PLS approach, the results of the RGCCA are correlations between latent variables and between manifest variables and their associated latent variables (there is no regression at the end of the algorithm).

The RGCCA is based on a simple iterative algorithm similar to that of the PLS approach which is as follows:

1 - Initialization of the outer weights in the same way as in the PLSPM algorithm.

2 - Standardization of the outer weights using the tau parameter:

$$w_j^0 = \left[\left(w_j^0\right)^T \left[\tau_j I + (1-\tau_j) \frac{1}{n} X_j^T X_j \right]^{-1} w_j^0 \right]^{-1/2} \left[\tau_j I + (1-\tau_j) \frac{1}{n} X_j^T X_j \right]^{-1} w_j^0$$

3 - Computation of the internal components of each latent variable depending on the scheme used (the inner schemes are the same as in PLSPM).

$$z_j^s = \sum_{k<j} c_{jk} e_{jk} X_k w_k^{s+1} + \sum_{k>j} c_{jk} e_{jk} X_k w_k^s$$

With e_{jk} being the inner weight and $c_{jk} = 1$ if the latent variables j and k are related.

4 - Outer weights are updated:

$$w_j^{s+1} = \left[\left(z_j^s\right)^T X_j \left[\tau_j I + (1-\tau_j)\frac{1}{n} X_j^T X_j\right]^{-1} X_j^T w_j^s \right]^{-1/2} \left[\tau_j I + (1-\tau_j)\frac{1}{n} X_j^T X_j\right]^{-1} X_j^T w_j^s$$

5 - We repeat steps 3 and 4 until convergence of the algorithm.

Once the algorithm has converged, we obtain results that optimize specific functions depending on the choice of the tau parameter.

Tau is a parameter that has to be set for each latent variable. It enables you to adjust the "mode" associated to the latent variable. If tau = 0, then we will be in the case of mode B and the results of PLSPM and RGCCA are similar. When tau = 1, we find ourselves in the new mode A (as stated by M. Tenenhaus). This mode is close to the mode A of PLSPM while optimizing a given function. When tau varies between 0 and 1, the latent variable mode stands in between mode A and mode B. For more details on RGCCA see Tenenhaus et al. (2011).

In the framework of RGCCA, XLSTAT-PLSPM allows to use the Ridge RGCCA mode. This mode search for the optimal tau parameter using the Schäfer and Stimmer (2005) formula reproduced in Tenenhaus et al. (2011).

The NIPALS algorithm

The roots of the PLS algorithm are in the NILES (Non linear Iterative LEast Squares estimation), which later became NIPALS (Non linear Iterative PArtial Least Squares), algorithm for Principal Component Analysis (Wold, 1966). We now remind the original algorithm of H. Wold and show how it can be included in the PLS-PM framework. The interests of the NIPALS algorithm are double as it shows: how PLS handles missing data and how to extend the PLS approach to more than one dimension.

The original NIPALS algorithm is used to run a PCA in presence of missing data. This original algorithm can be slightly modified to go into the PLS framework by standardizing the principal components. Once this is done, the final step of the NIPALS algorithm is exactly the Mode A of the PLS approach when only one block of data is available. This means that PLS-PM can actually yield the first-order results of a PCA whenever it is applied to a block of reflective manifest variables.

The other dimensions are obtained by working on the residuals of X on the previous standardized principal components.

The PLS approach for two sets of variables

PLS Path Modeling can be also used so as to find the main data analysis methods to relate two sets of variables. Table 1 shows the complete equivalence between PLS Path Modeling of two data tables and four classical multivariate analysis methods. In this table, the use of the deflation operation for the research of higher dimension components is mentioned.

Table 1: Equivalence between the PLS algorithm applied to two blocks of variables X_1 and X_2 and various method

--	Canonical Correlation analysis	Inter-Battery Factor Analysis	PLS Regression of X2 on X1	Redundancy Analysis of X2 with respect to X1
Computation mode for X1	B (deflation)	A (deflation)	A (deflation)	B (deflation)
Computation mode for X2	B (deflation)	A (deflation)	A (no deflation)	A (no deflation)

The analytical demonstration of the above mentioned results can be found in Tenenhaus et al., 2005.

TThe PLS approach for J sets of variables

The various options of PLS Path Modeling (Modes A or B for outer estimation; centroid, factorial or path weighting schemes for inner estimation) allow to find also many methods for multiple tables analysis: Generalized Canonical Analysis (the Horst's one (1961) and the Carroll's one (1968)), Multiple Factor Analysis (Escofier & Pagès, 1994), Lohmöller's split principal component analysis (1989), Horst's maximum variance algorithm (1965).

The links between PLS and these methods have been studied on practical examples in Guinot, Latreille and Tenenhaus (2001) and in Pagès and Tenenhaus (2001).

Let us consider a situation where J blocks of variables $X_1, \ldots, X_j$ are observed on the same set of statistical units. For estimating these latent variables ξ_j, Wold (1982) has proposed the hierarchical model defined as follows:

- A new block X is constructed by merging the J blocks $X_1, \ldots, X_j$ into a super block.
- The super block X is summarized by one latent variable ξ.
- A path model connects each exogenous LV ξ_j to the endogenous LV ξ.

An arrow scheme describing a hierarchical model for three blocks of variables is shown in Figure 1.

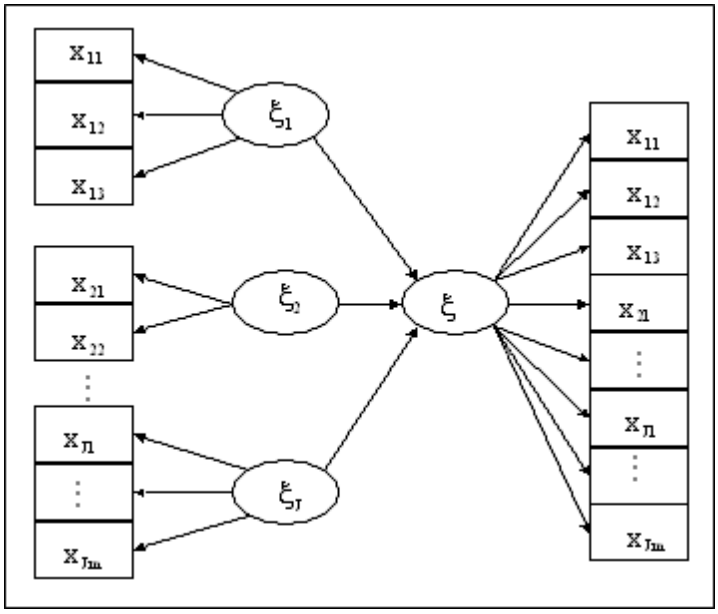


Figure 1: A hierarchical model for a PLS analysis of J blocks of variables.

Table 2 summarizes the links between Hierarchical PLS-PM and several multiple table analysis organized with respect to the choice of the outer estimation mode (A or B) and of the inner estimation scheme (Centroid, Factorial or Path Weighting).

Mode of calculation for the outer estimation	Centroid scheme	Factorial Scheme	Structural scheme
A	PLS Horst's generalized canonical correlation analysis	PLS Carroll's generalized canonical correlation analysis	Lohmöller's split principal component analysis / Horst's maximum variance algorithm /Excofier and Pagès Multiple Factor Analysis
B	Horst's generalized canonical correlation analysis (SUMCOR criterion)	Carroll's generalized canonical correlation analysis	

Table 2: PLS Path modeling and Multiple Table Analysis

In the methods described in Table 2, the higher dimension components are obtained by re-running the PLS model after deflation of the X-block.

It is also possible to obtain higher dimension orthogonal components on some X_j-blocks (or on all of them). The hierarchical PLS model is re-run on the selected deflated X_j-blocks.

The orthogonality control for higher dimension components is a tremendous advantage of the PLS approach (see Tenenhaus (2004) for more details and an example of application).

Finally, PLS Path Modeling may be meant as a general framework for the analysis of multiple tables. It is demonstrated that this approach recovers usual data analysis methods in this

context but it also allows for new methods to be developed when choosing different mixtures of estimation modes and schemes in the two steps of the algorithm (internal and external estimation of the latent variables) as well as different orthogonality constraints. Therefore, we can state that PLS Path Modeling provides a very flexible environment for the study of a multi-block structure of observed variables by means of structural relationships between latent variables. Such a general and flexible framework also enriches the data analysis methods with non-parametric validation procedures (such as bootstrap, jackknife and blindfolding) for the estimated parameters and fit indices for the different blocks that are more classical in a modeling approach than in data analysis.

Multigroup comparison tests in PLS path modeling

Two tests in order to compare parameters between groups are included in XLSTAT-PLSPM: An adapted t test based on bootstrap standard errors and a permutation test.

The multigroup t test:

Wynne Chin was the first to use this test to compare path coefficients. This test uses the estimates obtained from the bootstrap sampling in a parametric sense via t-tests. You make a parametric assumption and take the standard errors for the structural paths provided by the bootstrap samples, and then calculate the t-test for the difference in path coefficients between groups.

$$t = \frac{\left|\beta_{ij}^{G_1} - \beta_{ij}^{G_2}\right|}{\sqrt{\frac{(n_1-1)^2}{n_1+n_2-2}SE_{G_1}^2 + \frac{(n_2-1)^2}{n_1+n_2-2}SE_{G_2}^2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where n_1 and n_2 are the sizes of the groups, and $SE_{G_i}^2$ is the variance of coefficient β_{ij} obtained using the bootstrap sampling. This would follow a Student distribution with $n_1 + n_2 - 2$ degrees of freedom. This approach works reasonably well if the two samples are not far from normality and if the two variances are not too different.

The permutation tests:

Permutation tests offer a nonparametric alternative to t tests that fits well to PLS path modeling. They have been used together with PLS Path Modeling in Chin (2008) and Jakobowicz (2007). The principle is simple:

Select a statistic S. In the case of PLS Path Modeling, we take the absolute value of the difference on a parameter between two groups of observations.

Compute the value of this statistic on the two original samples associated to the groups à S_{obs} .

Randomly permute the elements of the two samples and compute the S statistic à S_{permi}. Repeat this step N_{perm} times (with N_{perm} very large).

The p-value is obtained with the following formula:

$$p - \text{valeur} = \frac{1}{N_{pem}+1}\sum_{i=1}^{N_{pen}} I\left(S_{obs} < S_{pem_i}\right)$$

Function $I(.)$ being 1 when $S_{obs} < S_{permi}$ and 0 otherwise.

Segmentation using the REBUS algorithm (Esposito Vinzi et al., 2008)

When heterogeneity at the units' level is present, it can be useful to obtain clusters of units which behave similarly on a defined model. The REBUS algorithm (Response Based procedure for detecting unit segments in PLS path modelling) offers a way to find clusters using a defined PLS model and residuals obtained using PLS Path Modeling. It is based on a simple algorithm:

1- Apply PLS Path Modeling on the entire dataset.

2- Compute residuals of the manifest and latent variables and apply an agglomerative hierarchical clustering on the residuals.

3- Apply PLS Path Modeling on each class of units using the same PLS model. You obtain as many sub-models as classes.

4- Compute residuals of the manifest and latent variables for each sub-model and compute the CM index for each unit and each class.

5- Allocate the units to the closest model (in term of CM).

6- Update the composition of the classes.

Repeat 3 to 6 until stable classes are obtained.

The CM index for unit i and class k is:

$$CM_{ik} = \sqrt{\frac{\sum_{j}\sum_{h=1}^{p_j}\left[e_{ihjk}^2 / \operatorname{Com}\left(\xi_{jk}, \mathbf{x}_{hj}\right)\right]}{\frac{\sum_{i=1}^{N}\sum_{j}\sum_{h=1}^{p_j}\left[e_{ihjk}^2 / \operatorname{Com}(\xi_{jk}, \mathbf{x}_{hj})\right]}{N-m_k-1}} \times \frac{\sum_{j^*=1}^{J^*}\left[f_{ij^*k}^2 / R^2\left(\xi_{j^*}, \xi_{j:\xi_j \to \xi_{j^*}}\right)\right]}{\frac{\sum_{i=1}^{N}\sum_{j^*=1}^{J^*}\left[f_{ij^*k}^2 / R^2\left(\xi_{j^*}, \xi_{j:\xi_j \to \xi_{j^*}}\right)\right]}{N-m_k-1}}}$$

Where Com() is the communality index, e_{ihjk} is the measurement model residual for the i-th unit in the k-th latent class, corresponding to the h-th manifest variable in the j-th block, i.e. the communality residual; f_{ij^*k} is the structural model residual for the i-th unit in the k-th latent class, corresponding to the j^*-th endogenous block; N is the number of units; m_k is the number of dimensions. Since all blocks are supposed to be reflective, there will always be one dimension (latent variable) per block.

XLSTAT-PLSPM offers some options to apply the REBUS method:

- Number of classes: The user can choose between automatic where the number of classes is set during the cluster analysis and manual where the number of classes has to be entered manually.
- Convergence: The algorithm converges when a fixed percentage of the units are staying in the same class from one iteration to another (the default value is 95%).

The REBUS method can only be applied if all blocks are reflective (mode A) and if no groups are selected.

The global quality index of the model allows to judge the quality of the segmentation. This index is equivalent to the GoF when a single class is used. When several classes are obtained, it is obtained as follows:

$$GQI = \sqrt{\sum_{k=1}^{K} \frac{n_k}{N} \left[\frac{1}{\sum P_q} \sum_{q} \sum_{p=1}^{P_q} \left(1 - \frac{\sum_{i=1}^{n_k} e_{ipqj}^2}{\sum_{i=1}^{n_k} \left(x_{ipq} - \overline{x}_{pqk} \right)^2} \right) \right]} \times \sqrt{\sum_{k=1}^{K} \frac{n_k}{N} \left[\frac{1}{J} \sum_{j=1}^{J} \left(1 - \frac{\sum_{i=1}^{n_k} f_{ijk}^2}{\sum_{i=1}^{n_k} \left(\hat{\xi}_{ipq} - \overline{\hat{\xi}}_{pqk} \right)^2} \right) \right]}$$

with n_k being the size of class k, P_q being the number of manifest variable associated to the latent variable q. e^2 is the residual obtained from the manifest variables and f^2 is the residual obtained from the scores of the latent variables.

In addition to the GQI, XLSTAT-PLSPM displays the overall improvement of the GQI and the decomposition of the GQI based on the outer model and on the inner model.

The Marketing display for the analysis of consumer satisfaction

In addition to the Classic and Expert displays, XLSTAT-PLSPM Proposes a **Marketing** display that provides the user with an interface that is specially designed for marketing and satisfaction analysis.

This display sets default values to numerous parameters and proposes a simplified way to deal with variables: the scales of manifest and latent variables are selected from the beginning (see the **Fitting the Model** paragraph).

By default, the PLS-Path Modeling method, the path weighting (structural) scheme as well as OLS regressions are used. Initial weights are obtained with the values of the first principal component analysis eigenvector and one single dimension is displayed.

The segmentation option allows to carry out REBUS analysis to extract classes of homogeneous observations and multigroup tests can be defined in the **options** tab.

This display also proposes charts and simulation tables that helps evaluating the impact of changing variables (manifest and latent) on a target latent variable.

Model loading from the model library

XLSTAT-PLSPM lets you easily load existing structural models. Use the **load model** icon and choose the model to load among inside the model library.

The loaded model replaces the model that is already open but keeps your data and your results. However, the manifest and latent variables association operation should be done in addition.

Below is a list of available models in the library (.ppmxmod files):

- 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)
- Simplified ECSI (excludes the *complaints* variable)

Projects

XLSTAT-PLSPM projects are special Excel workbook templates. When you create a new project, its default name starts with PLSPMBook. You can then save it to the name you want, but make sure you use the "Save" or "Save as" command of the XLSTAT-PLSPM toolbar to save it in the folder dedicated to the PLSPM projects using the *.ppm extension with Excel 2003 and prior, and *.ppmx with Excel 2007 and later.

A raw XLSTAT-PLSPM project contains two sheets that cannot be removed:

- D1: This sheet is empty and you need to add all the input data that you want to use into that worksheet.

- PLSPMGraph: This sheet is blank and is used to design the model. When you select this sheet, the "Path modeling" toolbar is displayed. It is made invisible when you leave that sheet.

Once a model has been designed, you can run the optimization. Results sheets are then added after the PLSPMGraph sheet.

It is possible to record a model before adding new variables and to reload it later (see the Toolbars section for additional information).

Each time you create a new project, XLSTAT asks you what display you wish to use. You may choose among 3 options depending on your objectives and your expertise.

It is also possible to load saved models (see the **toolbars** paragraph for more details).

Options

To display the options dialog box, click the  button of the "XLSTAT-PLSPM" toolbar. Use this dialog box to define the general options of the XLSTAT-PLSPM module.

General tab:

Display: Because PLS Path Modeling and XLSTAT-PLSPM are complex, three display styles are available. The classical one including all commonly used options and the expert one lets you use more advanced features. Finally, the Marketing display lets you obtain a configuration adapted to marketing and satisfaction analysis.

Path for the XLSTAT-PLSPM projects: This path can be modified if and only if you have administrator rights on the machine. You can then modify the folder where the user's files are saved by clicking the […] button that will display a box where you can select the appropriate folder. The folder must be accessible for reading and writing to all types of users.

Format tab:

Use these options to set the format of the various objects that are displayed on the PLSPMGraph sheet:

- **Latent variables**: You can define the color and the size of the border line of the ellipses that represent the latent variables, as well as the color of the background, and the color and the size of the font.
- **Manifest variables**: You can define the color and the size of the border line of the rectangles that represent the manifest variables, as well as the color of the background, and the color and the size of the font.
- **Arrows (MV-LV)**: You can define the color and the size of the arrows between the manifest and the latent variables.
- **Arrows (LV-LV)**: You can define the color and the size of the arrows between two latent variables.

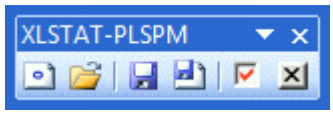
Note 1: So that changes are taken into account once you click the OK button, you need to optimize the display by clicking on the  button.

Note 2: these options do not prevent you from changing the format of one or more objects on the PLSPMGraph sheet. Using the drawing toolbar of Excel, you can change the fill, the borders of the objects.

Toolbars

XLSTAT-PLSPM has two toolbars, "XLSTAT-PLSPM" and "Path modeling".

The "XLSTAT-PLSPM" toolbar can be displayed by clicking the XLSTAT-PLSPM icon in the XLSTAT toolbar.



Click this icon to open a new PLSPM project (see Projects for more details).

Click this icon to open an existing PLSPM project.

Click this icon to save the current PLSPM project. This icon is only active if changes have been made in the project.

Click this icon to save the project in a new folder or under a new name.

Click this icon to display the XLSTAT-PLSPM options dialog box.

Click this icon if you want to continue using XLSTAT but not XLSTAT-PLSPM. This allows to free some memory space.

Excel 2003 and lower:

The second toolbar, "Path modeling" is only visible when you are on the PLSPMGraph sheet of a PLSPM project.

Click this icon to add latent variables. If you double click this icon, you can add several latent variables in a row without having to click this button each time.

Click this icon to add an arrow between two latent variables. If you double click this icon, you can add several arrows in a row without having to click this button each time. When adding an arrow, select first the latent variable that will be at the beginning of the arrow by clicking on it, then leave the mouse button down, and drag the cursor till the latent variable that will be at the end of the arrow.

Click this icon to hide the manifest variables. If a latent variable is selected when you click this icon, it will only hide the corresponding manifest variables.

Click this icon to display the manifest variables. If a latent variable is selected when you click this icon, it will only show the corresponding manifest variables.

Click this icon to optimize the display.

Click this icon to define groups. Once groups are defined, a list with the group names is displayed on the PLSPMGraph sheet, and the icon becomes ; click this icon to remove the groups definition.

Click this icon to archive the current model in the project.

Click this icon to load a previously saved model.

Click this icon to delete one or more previously saved models.

Unprotected/Protected(1)/Protected(2): The first option allows the user to modify the model and the position of the objects. The second option allows the user to modify the position of the objects. The third option does not allow the user to move the objects or to delete them.

Click this icon to completely remove all the objects on the PLSPMGraph sheet.

Click this icon to display the results of the model, if the latter has already been fitted. If the results are already displayed, the following icon is displayed ; click it to hide the results.

Click this icon to display a dialog box that allows you to choose which results should be displayed or not.

Click this icon to start the optimization of the model, and then display the results on both the PLSPMGraph sheet, and on the results sheet.

Excel 2007 and later:

The second toolbar, "Path modeling" is included in the PLSPMGraph sheet of a PLSPM project.

Click this icon to add latent variables. If you double click this icon, you can add several latent variables in a row without having to click this button each time.

Click this icon to define the manifest variables, after selecting the latent variable on the diagram. You can also use the keyboard shortcut Ctrl+M.

Click this icon to hide the manifest variables. If a latent variable is selected when you click this icon, it will only hide the corresponding manifest variables.

Click this icon to display the manifest variables. If a latent variable is selected when you click this icon, it will only show the corresponding manifest variables.

Click this icon to optimize the display.

 Click this icon to define groups. Once groups are defined, a list with the group names is displayed on the PLSPMGraph sheet, and the icon becomes  ; click this icon to remove the groups definition.

 Click this icon to archive the current model in the project.

 Click this icon to load a previously saved model or to load a model from a model library (see the description section).

 Click this icon to delete one or more previously saved models.

Unprotected/Protected(1)/Protected(2): The first option allows the user to modify the model and the position of the objects. The second option allows the user to modify the position of the objects. The third option does not allow the user to move the objects or to delete them.

 To add a link between two latent variables: When adding a link, select first the latent variable that will be at the beginning of the arrow by clicking on it, then select the latent variable that will be at the end of the arrow. Click this icon when both variables are selected. You can also use the keyboard shortcut Ctrl+L. Use the keyboard shortcut Ctrl+R to reverse the direction of the link.

 Click this icon to transform all the links into double direction links. This can also be obtained with the keyboard shortcut Ctrl+D.

 Click this icon to change the position of the manifest variables. You can also use the keyboard shortcut Ctrl+O.

 Click this icon to rename a latent variable.

 Click this icon to completely remove all the objects on the PLSPMGraph sheet.

 Click this icon to display the results of the model, if the latter has already been fitted. If the results are already displayed, the following icon is displayed  ; click it to hide the results.

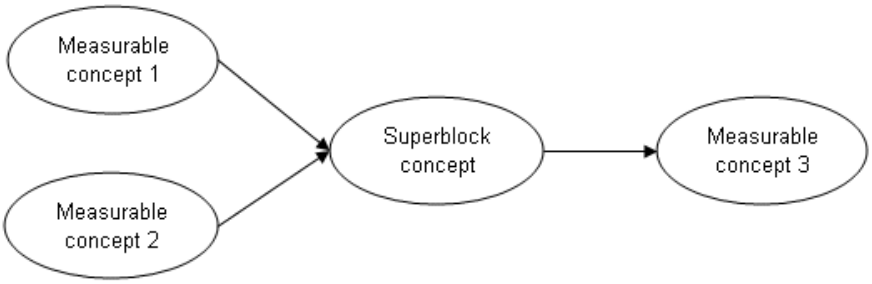
 Click this icon to display a dialog box that allows you to choose which results should be displayed or not.

 Click this icon to start the optimization of the model, and then display the results on both the PLSPMGraph sheet, and on the results sheet.

To delete an arrow between two latent variables, you can click on the arrow, then use Ctrl+Delete.

Adding manifest variables

Once one or more latent variables have been added on the PLSPMGraph document using the appropriate icon of the "Path modeling" toolbar, you can define the manifest variables that correspond to these variables (with Excel 2003 and lower, make a double click on the latent variable, with Excel 2007 and later, click the Manifest variables icon). A latent variable must have manifest variables, even if it is a superblock variable (a variable that is not directly related to manifest variables but to latent variables with arrows going from the latent variables to the superblock variable – the superblock variable inherits the manifest variables of the constitutive latent variables).



For a superblock, you need to add all the manifest variables of the parent latent variables. This is made easy with the XLSTAT interface.

To add manifest variables, you can

Double-click the latent variable;

Click the right button of the mouse, and select "Add manifest variables".

These actions lead to the display of a dialog box which options are:

General tab:

Name of the latent variable: Enter the name of the latent variable.

Manifest variables: Select on the D1 sheet the data that correspond to the manifest variables. The input variables can be either quantitative or qualitative.

- **Quantitative**: Activate this option if you want to use quantitative variables and then select these variables.
- **Qualitative**: Activate this option if you want to use qualitative variables and then select these variables.

Variable labels: Activate this option if the first row of the data selections includes a header.

Position: Select the position where the manifest variables should be positioned relatively to the latent variable.

Mode: Select the mode option that determines how the latent variable is constructed from the manifest variables. The available options are "**Mode A** " (reflective way, arrows go from the latent variable to the manifest variables), "**Mode B** " (formative way, arrows go from the manifest variables to the latent variable), "**Centroid** ", "**PCA** ", "**PLS** ", and mode " **MIMIC** " (a mixture of Mode A and Mode B). If Mode MIMIC is selected, you need to select a column with one row per manifest variable (and a header if the "Variable labels" option is checked), with As for the variables with Mode A, and Bs for the variables with mode B. See the description section for more information on the modes. The "**Automatic** " mode is only available for superblocks. It allows to make that the mode for each manifest variable corresponds to its mode in the latent variable that is used to create the superblock. "Centroid", "PCA", "PLS", and "Automatic" modes are only available in the expert display. The **RGCCA** mode allows to enter the value of the tau parameter and the Ridge RGCCA mode automatically optimize the value of the tau parameter. These two modes can only be applied with the RGCCA method (see the description section).

Deflation (expert display): Select the deflation mode. The deflation is used when computing the model on the second and above dimensions of the model.

- **No deflation**: Whatever the dimension, the scores of the latent variable remain constant.
- **External**: For the successive dimensions, the residuals are computed from the outer model.

Dimension (expert display)**:** You can choose the number of dimensions to be studied.

Invert sign: Activate this option if you want to reverse the sign of the latent variable. This option is useful if you notice that the influence of a latent variable is the opposite of what it should be.

Supberblock (expert display): You can activate this option only if latent variables have already been created, and if manifest variables were selected for the latter. The list displays the latent variables for which manifest variables have already been defined. The superblock tab appears. You can then select the latent variables that are used to build the superblock variable.

Interaction (expert display): You can activate this option only if latent variables have already been created. An interaction latent variable is the product of two latent variables that have the same successor. The interaction variable will have the same successor as the two variables that were used to create it. The interaction tab appears.

Superblock tab:

The list of all latent variables is displayed. Select latent variables to be included in the superblock.

Interaction tab:

Generating latent variable: Select two of the latent variables explaining the latent variable to which the interaction variable is connected.

Treatment of the manifest variables: Select which transformation of the manifest variable prior to the product you wish to apply. Three options are available: raw manifest variables, standardized manifest variables and mean centered manifest variables.

Options (PLS) tab (expert display):

Options in the structural model (PLS):

Stop conditions:

- **Automatic**: Activate this option so that XLSTAT automatically determines the number of components to keep.
- **Max components**: Activate this option to set the pour fixer le maximum number of components to take into account in the model.

Options for PLS regression in the **measurement model** (only active, if the "PLS" mode is selected):

Stop conditions:

- **Automatic**: Activate this option so that XLSTAT automatically determines the number of components to keep.
- **Max components**: Activate this option to set the pour fixer le maximum number of components to take into account in the model.

Defining groups

If a qualitative variable is available and if you believe that the model could be different for the various categories of that variable, you may use it to define groups.

To define groups, go to the "PLSPMGraph" sheet, then click the appropriate icon. This displays a "Groups" dialog box, which entries are:

Groups: Select on the D1 sheet the data that correspond to the qualitative variable that indicates to which group each observation belongs.

Column label: Activate this option if the first row of the selection corresponds to a header.

Sort alphabetically: Activate this option if you want that XLSTAT sorts alphabetically the names of the groups (the categories of the selected qualitative variable). If this option is not activated, the categories are listed in their order or appearance.

Once you click **OK**, a list is added at the top right corner of the PLSPMGraph sheet. Once the model has been computed, you can use this list to display the results of the group you want on the PLSPMGraph sheet. The results of the model that corresponds to each group are also displayed on distinct sheets.

Note: if you want to remove the group information, click the appropriate button of the "Path modeling" toolbar.

Once groups are defined, multigroup tests can be performed.

Fitting the model

Once you have designed the model on the PLSPMGraph sheet, and once the manifest variables have been defined for each latent variable, you can click the appropriate icon of the "Path modeling" toolbar to display the "Run" dialog box that lets you define additional options before fitting the model.

General tab:

Treatment of the manifest variables (classic and expert displays only): Choose if and how the manifest variables should be transformed.

- **Standardized, weights on standardized MV**: The manifest variables are standardized before fitting the model, and the corresponding outer weights are estimated.
- **Standardized, weights on raw MV**: The manifest variables are standardized before fitting the model, and the outer weights are estimated for the raw variables.
- **Reduced, weights on raw MV**: The manifest variables are reduced (divided by the standard deviation) before fitting the model, and the corresponding outer weights are estimated.
- **Raw MV**: The manifest variables are not transformed.

Initial weights (classic and expert displays only) : Choose which initial value should be used for outer weight initialization.

Values of the first eigenvector. The initial values are the values associated to the first eigenvector.

Signs of the coordinates of the first eigenvector. Instead of taking the values of the first eigenvector only take the sign.

-1 for max of first eigenvector, else +1.

-1 for min of first eigenvector, else +1.

+1 for variable with max variance, else 0.

-1 for the last manifest variable, else +1.

Weights: Activate this option if the observations are weighted. If you do not activate this option, the weights will be considered as 1. Weights must be greater than or equal to 0. If a column header has been selected, check that the "Variable labels" option is activated.

Method (classic and expert displays only): Choose the method to be used. You can choose between PLSPM, GSCA and RGCCA.

REBUS (expert display only) : Activate this option if you want to apply the REBUS algorithm on your model. When you activate this option the REBUS tab is reachable. When using REBUS, all blocks of manifest variables have to be reflective (mode A), and manifest and latent variables have to be standardized. The NIPALS and Lohmöller options are not available for the treatment of missing data.

RGCCA (expert display only): Activate this option if you want to apply the RGCCA. All the blocks have to be defined with either mode A, mode B or the RGCCA mode.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Variable labels: Activate this option if the first row of the data selections includes a header.

Observation labels: Activate this option if observations labels are available. Then select the corresponding data. If the "Variable labels" option is activated you need to include a header in the selection. If this option is not activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 …).

Scale of the manifest variables (Marketing display): allows to choose the scale on which the manifest variables are presented. Two options are available.

- Automatic: manifest variables are standardized.
- Select: enter the minimum and maximum values of the manifest variables (uniform scale) in order to work on the original variables.

Scale of the latent variables (Marketing display):

- 0-100 scale: latent variables scores are given on a 0-100 scale.
- MV scale: latent variables scores are given on the same scale as the manifest variables.

Options tab:

Internal estimation (classic and expert displays only): Select the internal estimation method (see the description section for additional details).

- **Structural**: The inner weights are equal to the correlation between the latent variables when estimating an explanatory (predecessor) latent variable. Otherwise they are equal to the OLS regression coefficients.
- **Factorial**: The inner weights are equal to the correlation between the latent variables.

- **Centroid**: The inner weights are equal to the sign of the correlation between the latent variables.
- **PLS**: The inner weights are equal to the correlation between the latent variables when estimating an explanatory (predecessor) latent variable. Otherwise they are equal to the PLS regression coefficients.

Regression (expert display only): Select the regression method that is used to estimate path coefficients

- **OLS**: Ordinary Least Squares regression.
- **PLS**: Partial Least Squares regression.

Dimensions (expert display only): Enter the number of dimensions up to which the model should be computed.

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 100.
- **Convergence**: Enter the value of the difference of criterion between two steps which, when reached, means that the algorithm is considered to have converged. Default value: 0.0001.

Confidence intervals: Activate this option to compute the confidence intervals. Then choose the method used to compute the intervals:

- **Bootstrap**: Activate this option to use a bootstrap method. Then enter the number of "Re-samples" generated to compute the bootstrap confidence intervals.
- **Jackknife**: Activate this option to use a jackknife method. Then enter the "Group size" that is used to generate the samples to compute the jackknife confidence intervals.

Confidence interval (%): Enter the size in % of the confidence intervals.

Resampled estimates (expert display only): Activate this option to display estimates of standardized loading and path coefficients for each sample being generated. Standard deviation and bounds of the confidence intervals for indirect effects are also displayed.

Model quality (classic and expert displays only):

- **Blindfolding**: Activate this option to check the model quality using the blindfolding approach (see the description section for additional details). Cross-validated values for redundancy and communality will be computed.

Segmentation (Marketing display): activate this option to look for classes of homogeneous observations with regards to the specified model. The REBUS algorithm is used. A truncation can also be applied.

Comparisons (Marketing display): if groups have been selected, permutation t-tests can be activated here.

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations that contain missing data.

Use NIPALS: Activate this option to use the NIPALS algorithm to handle missing data (see the description section for additional details).

Lohmöller: Activate this option to use the Lohmöller's procedure to handle missing data: pairwise deletion to compute sample means and standard deviation, and mean imputation to compute the scores. If PLS regression is used instead of OLS regression, then model is applied on the available data.

- **Use the ipsative mean**: Activate this option to use the mean of the latent variables to estimate missing data in the manifest variables.
- **Renormalize**: Activate this option to renormalize external weights for each observation when missing data have been found.

Note: In the case of standardized weights, the two options above lead to pairwise deletion to compute sample means and standard deviation, and mean imputation to compute the scores.

Estimate missing data: Activate this option to estimate missing data before starting the computations.

- **Mean or mode**: Activate this option to estimate missing data by using the mean (quantitative variables) or the mode (qualitative variables) of the corresponding variables.
- **Nearest neighbor**: Activate this option to estimate the missing data of an observation by searching for the nearest neighbor of the observation.

Multigroup tests tab (expert display):

If groups have been selected, this tab appears.

Multigroup t test: Activate this option to test equality between path coefficients from one group to another with a t test (the number of bootstrap sample is defined in the option tab).

- **Significance level (%)**: Enter the significance level for the t tests.

Permutation tests: Activate this option to test equality of parameters between two groups with a permutation test.

- **Number of permutations:** Enter the number of permutations.
- **Significance level (%)**: Enter the significance level for the t tests.
- **Path coefficients:** Activate this option to test the equality of the path coefficients.
- **Standardized loadings:** Activate this option to test the equality of the standardized loadings.

Model quality: Activate this option to test the equality of the quality indexes (communalities, redundancies, and GoF).

REBUS tab (expert display):

If the REBUS option is activated, this tab appears.

Truncation: Activate this option is you want XLSTAT to **automatically** define the truncation level, and therefore the number of classes to retain, or if you want to define the **number of classes** to create, or the **level** at which the dendrogram shall be truncated.

Stop conditions:

- **Iterations**: Enter the maximum number of iterations for the REBUS algorithm. The calculations are stopped when the maximum number of iterations has been exceeded. Default value: 100.
- **Threshold (%)**: Enter the percentage of stable units necessary to stop the algorithm. Default value: 95.

Dendrogram:

- **Horizontal**: Choose this option to display a horizontal dendrogram.
- **Vertical**: Choose this option to display a vertical dendrogram.
- **Labels**: Activate this option to display object labels (full dendrogram) or classes (truncated dendrogram) on the dendrogram.
- **Colors**: Activate this option to use colors to represent the different groups on the full dendrogram.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the selected variables.

Model: Activate this option to display the model specifications.

Correlations (classic and expert designs only): Activate this option to display the correlation matrix.

- **Test significance**: Activate this option to test the significance of the correlations.
- **Significance level (%)**: Enter the significance level for the above tests.

Outliers' analysis (expert design only): Activate this option to display the DmodX and DModY table when PLS regression is selected.

MV after deflation (expert design only): Activate this option to display the manifest variables after deflation when more than one dimension has been selected.

Variables/Factors correlations: Activate this option to display correlations between factors and variables.

Inner model: Activate this option to display the results that correspond to the inner model.

Outer model: Activate this option to display the results that correspond to the outer model.

R^2 and communalities: Activate this option to display the R2 of the latent variables from the structural model and the communalities of the manifest variables.

Model quality: Activate this option to display the results of the blindfolding procedure.

Latent variable scores (classic and expert designs only):

- **Standardized**: Activate this option to compute and display standardized factor scores.
- **Using normalized weights**: Activate this option to display factor scores computed with normalized weights.
- **Standardized > 0-100**: Activate this option to compute standardized scores, and then transform and display the latter on a 0-100 scale.
- **Using normalized weights > 0-100**: Activate this option to compute factor scores using normalized weights, and then transform and display the factor scores on a 0-100 scale.

Simulation table (Marketing display): activate this option to display simulation tables that let you visualize the effects of the modification of a manifest or a latent variable on a target latent variable.

- LV to explain: select the target latent variable to explain. You should select an endogeneous latent variable.
- Scale of changes: select the scale of changes (percent or number of points). Once this option is configured, you will be able to enter the minimum, the maximum as well as the

change step to obtain the range of values to test.

IPMA (Marketing display): Activate this option if you wish to display the tables based on IPMA (Importance Perform Analysis).

Charts tab:

Coefficients plot: activate this option to display normalized coefficients of the internal model.

IPMA chart: activate this option to display the IPMA charts.

Simulation plot (Manifest variables) (Marketing display): activate this option to display simulation plots to investigate the effects of modifying manifest variables on the score of the target latent variable.

Simulation plot (Latent variables) (Marketing display): activate this option to display simulation plots to investigate the effects of modifying latent variables on the score of the target latent variable.

Results options

Many results can be displayed on the PLSPMGraph sheet, once the model has been fitted. It is recommended to select only a few items in order to keep the results easy to read. To display the options dialog box, click the results icon of the "Path modeling" toolbar.

Latent variables tab:

These options allow defining which results are displayed below the latent variables.

- **Mean:** Activate this option to display the mean of the latent variable.
- **Mean (Bootstrap):** Activate this option to display the mean of the latent variable computed using a bootstrap procedure.
- **Confidence interval:** Activate this option to display the confidence interval for the mean.
- **R^2:** Activate this option to display the R-square between the latent variable and its manifest variables.
- **Adjusted R^2:** Activate this option to display the adjusted R-square between the latent variable and its manifest variables.
- **R^2 (Boot/Jack):** Activate this option to display the R-square between the latent variable and its manifest variables, computed using a bootstrap or jackknife procedure.
- **R^2 (conf. int.):** Activate this option to display the confidence interval on the R-square between the latent variable and its manifest variables, computed using a bootstrap or jackknife procedure.
- **Communality:** Activate this option to display the communality between the latent variable and its manifest variables.
- **Redundancy:** Activate this option to display the redundancy between the latent variable and its manifest variables.
- **Communality (Blindfolding):** Activate this option to display the communality between the latent variable and its manifest variables, computed using the blindfolding procedure.
- **Redundancy (Blindfolding):** Activate this option to display the redundancy between the latent variable and its manifest variables, computed using the blindfolding procedure.
- **D.G. rho:** Activate this option to display the Dillon-Goldstein's rho coefficient.
- **Cronbach's alpha:** Activate this option to display the Cronbach's alpha.
- **Std. deviation (Scores):** Activate this option to display the standard deviation of the estimated latent variables' scores

Arrows (Latent variables) tab:

These options allow to define which results are displayed on the arrows that relate the latent variables.

- **Correlation:** Activate this option to display the correlation coefficient between the two latent variables.
- **Contribution:** Activate this option to display the contribution of the latent variables to the R2.
- **Path coefficient:** Activate this option to display the regression coefficient that corresponds to the regression of the latent variable that is at the end of the arrow (dependent) by the latent variable at the beginning of the arrow (predecessor or explanatory).
- **Path coefficient (B/J):** Activate this option to display the regression coefficient that corresponds to the regression of the latent variable that is at the end of the arrow (dependent) by the latent variable at the beginning of the arrow (predecessor or explanatory), computed using a bootstrap or jackknife procedure.
- **Standard deviation:** Activate this option to display the standard deviation of the path coefficient.
- **Confidence interval:** Activate this option to display the confidence interval for the path coefficient.
- **Std. coeff.:** Activate this option to display standardized coefficients.
- **Student's t:** Activate this option to display the value of the Student's t.
- **Partial correlations:** Activate this option to display partial correlations between latent variables.
- **Pr > |t|:** Activate this option to display the p-value that corresponds to the Student's t.
- **Arrows thickness depends on:** The thickness of the arrows can be related to:
- The p-value of the Student's t (the lower the value, the thicker the arrow).
- The correlation (the higher the absolute value, the thicker the arrow; blue arrows correspond to negative values, red arrows to positive values).
- The contribution (the higher the value, the thicker the arrow).

Arrows (Manifest variables) tab:

These options allow to define which results are displayed on the arrows that relate the latent variables.

- **Weight:** Activate this option to display the weight.

- **Weight (Bootstrap):** Activate this option to display the weight computed using a bootstrap procedure.
- **Normalized weight:** Activate this option to display the normalized weight.
- **Standard deviation:** Activate this option to display the standard deviation of the weight.
- **Confidence interval:** Activate this option to display the confidence interval for the weight.
- **Correlation:** Activate this option to display the correlation coefficient between the manifest variable and the latent variable.
- **Correlation (Boot/Jack):** Activate this option to display the correlation coefficient between the manifest variable and the latent variable, computed using a bootstrap of jackknife procedure.
- **Correlation (std. deviation):** Activate this option to display the standard deviation of the correlation coefficient between the manifest variable and the latent variable, computed using a bootstrap of jackknife procedure.
- **Correlation (conf. interval):** Activate this option to display the confidence interval of the correlation coefficient between the manifest variable and the latent variable, computed using a bootstrap of jackknife procedure.
- **Communalities:** Activate this option to display the communality between the latent variable and the manifest variables.
- **Redundancy:** Activate this option to display the redundancy between the latent variable and the manifest variables.
- **Communality (Blindfolding):** Activate this option to display the communality between the latent variable and its manifest variables, computed using the blindfolding procedure.
- **Redundancy (Blindfolding):** Activate this option to display the redundancy between the latent variable and its manifest variables, computed using the blindfolding procedure.
- **Arrows thickness depends on:** The thickness of the arrows can be related to:
- The correlation (the higher the absolute value, the thicker the arrow; blue arrows correspond to negative values, red arrows to positive values).
- Normalized weights.

Results

The first results are general results which computation is done prior to fitting the path modeling model:

Summary statistics: This table displays for all the manifest variables, the number of observations, the number of missing values, the number of non- missing values, the minimum, the maximum, the mean and the standard deviation.

Model specification (measurement model): This table displays for each latent variable, the number of manifest variables, the mode, the type (a latent variable which never appears as a dependent variable is called exogenous), if its sign has been inverted, the number of computed dimension and the list of all associated manifest variables.

Model specification (structural model): This square matrix shows on its lower triangular part if there is an arrow that goes from the column variable to the row variable.

Composite reliability: This table allows to check the dimensionality of the blocks. For each latent variable, a PCA is run on the covariance or correlation matrix of the manifest variables in order to determine the dimensionality. The Cronbach's alpha, the Dillon-Goldstein's rho, the critical eigenvalue (that can be compared to the eigenvalues obtained from the PCA) and the condition number are displayed to facilitate the determining of the dimensionality.

Variables/Factors correlations (Latent variable X / Dimension Y): These tables display for each latent variable the correlations between the manifest variables and the factors extracted from the PCA. When a block is not unidimensional, these correlations allow to identify how the corresponding manifest variables can be split into unidimensional blocks.

The results that follow are obtained once the path modeling model has been fitted:

Goodness of fit index (Dimension Y): This table displays the goodness of fit index (GoF) computed using bootstrap or not and its confidence interval for

- Absolute: Value of the GoF index.
- Relative: Value of the relative GoF index obtained by dividing the absolute value by its maximum value achievable for the analyzed dataset.
- Outer model: Component of the GoF index based on the communalities (performance of the measurement model).
- Inner model: Component of the GoF index based on the R2 of the endogenous latent variables (performance of the structural model).

Cross-loadings (Monofactorial manifest variables / Dimension Y): This table allows to check whether a given manifest variable is really monofactorial, i.e. mostly related to its latent

variable or if it is also related to other variables. Ideally, if the model has been well specified, it should appear as being mostly related to its latent variable.

Outer model (Dimension Y):

- **Weights (Dimension Y)**: Coefficients of each manifest variable in the linear combination used to estimate the latent variable scores.
- **Standardized loadings (Dimension Y)**: Correlations (standardized loadings) between each manifest variable and the corresponding latent variable. Loadings and location parameters are also displayed.

Inner model (Dimension Y):

- **R^2 (Latent variable X / Dimension Y)**: Value of the R2 index for the endogenous variables in the structural equations.
- **Path coefficients (Latent variable X / 1)**: Value of the regression coefficients in the structural model estimated on the standardized factor scores. The size effect (f^2) is also displayed.
- **Impact and contribution of the variables to Latent variable X (Dimension Y)**: Value of the path coefficients and the contributions (in percent) of the predecessor latent variables to the R2 index of the endogenous latent variables.

Bootstrap: Values of the standardized loadings and path coefficients for each generated sample.

Model assessment (Dimension Y): This table summarizes important results associated to the latent variables scores.

Correlations (Latent variables) / Dimension Y (Expert display): Correlation matrix obtained on the latent variable scores.

Partial correlations (Latent variables) / Dimension Y (Expert display): Partial correlation matrix obtained on the latent variable scores.

Direct effects (latent variable) / Dimension Y (Expert display): This table shows direct effect between connected latent variables.

Indirect effect (latent variable) / Dimension Y (Expert display): This table shows the indirect effects between not directly connected latent variables. If the resampled estimates option has been selected, the standard deviations and the bounds of the confidence intervals are also displayed.

Total effect (latent variable) / Dimension Y (Expert display): This table shows the total effect between latent variables. Total effect = direct effect + indirect effect.

Discriminant validity (Squared correlations < AVE) (Dimension Y): This table allows to check whether each latent variable is really representing a concept different from the other or if some latent variables are actually representing the same concept. In this table, the R2 index for any pair of latent variables shall be smaller than the mean communalities for both variables

which indicates that more variance is shared between each latent variable and its block of manifest variables than with another latent variable representing a different block of manifest variables.

IPMA (Importance Performance Matrix Analysis) tables and charts (Expert and Market displays): for each endogeneous latent variable, those tables gather importance and performance values of the latent variables. Importance is the total effect on the studied endogeneous latent variable. Performance is the score of the latent variable scaled between 0 and 100. Those indices are represented on charts.

Simulation tables and plots (Marketing display): those results can be used to understand the impact of the modification of a variable in the model on a target latent variable to explain.

- The first table gathers the most important latent variables for the prediction of the target latent variable to explain.
- The second table displays the most important manifest variables for the prediction of the target latent variable to explain.
- The following table and chart allow to visualize the impact of modifying a manifest variable on the target latent variable to explain.
- The following table and chart allow to visualize the impact of modifying a manifest variable on the score mean of the target latent variable to explain (the mean is displayed and not the change).
- The following table and chart allow to visualize the impact of modifying a latent variable on the target latent variable to explain.
- The following table and chart allow to visualize the impact of modifying a latent variable on the score mean of the target latent variable to explain (the mean is displayed and not the change).

Latent variable scores (Dimension Y):

- Mean / Latent variable scores (Dimension Y): Mean values of the individual factor scores.
- Summary statistics / Latent variable scores (Dimension Y): Descriptive statistics of the latent variable scores computed from the measurement model.
- Latent variable scores (Dimension Y): Individual latent variable scores estimated as a linear combination of the corresponding manifest variables.
- Summary statistics / Scores predicted using the structural model (Dimension Y) (expert display): Descriptive statistics of the latent variable scores computed from the structural model.
- Scores predicted using the structural model (Dimension Y) (expert display): Latent variable scores computed as the predicted values from the structural model equations.

Model assessment / Outer model (Blindfolding): Cross-validated values of the communalities obtained by means of the blindfolding procedure.

Model assessment / Inner model (Blindfolding): Cross-validated values of the redundancies obtained by means of the blindfolding procedure.

If groups are defined, some other outputs are available:

Worksheet PLSPM (Group): For each group, complete results are displayed in separated worksheets.

Worksheet PLSPM (Multigroup t test): For each path coefficient, results of the t test are summarized in a table. Each line represents a pair of groups.

- Difference: Absolute value of the parameter's difference between the groups.
- t (Observed value): Observed value of the t statistic.
- T (critical value): Critical value of the t statistic.
- DF: Number of degrees of freedom.
- p-value: p-value associated to the t test.
- Alpha: Significance level
- Significant: If yes, the difference between the parameters is significant. If not, the difference is not significant.

Worksheet PLSPM (Permutation test): For each type of parameter, results of the permutation test are summarized in a table.

- Difference: Absolute value of the parameter's difference between the groups.
- p-value: p-value associated with the permutation test.
- Alpha: Significance level.
- Significant: If yes, the difference between the parameters is significant. If not, the difference is not significant.

If the REBUS option is activated, some other outputs are available:

Worksheet REBUS : The dendogram obtained with the cluster analysis is displayed. For each observation, the class and the CM index is also displayed.

Worksheet PLSPM (Class): For each class, complete results are displayed in separated worksheets.

Example

A tutorial on how to use the XLSTAT-PLSPM module with Excel 2007 is available on the XLSTAT Help Center:

http://www.xlstat.com/demo-plspm2007.htm

A tutorial on how to use the XLSTAT-PLSPM module with Excel 2003 is available on the XLSTAT Help Center:

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

A tutorial on how to compare groups with XLSTAT-PLSPM is available on the XLSTAT Help Center:

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

A tutorial on how to use the REBUS method with XLSTAT-PLSPM in available on the XLSTAT Help Center:

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

References

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

Latent class clustering

This tool is part of the XLSTAT-LG module. Use this tool to classify cases into meaningful clusters (latent classes) that differ on one or more parameters from latent class (LC) Cluster models. LC Cluster models classify based on combinations of continuous and/or categorical (nominal or ordinal) variables.

In this section:

Description

This the latent class clustering feature of XLSTAT is part of the XLSTAT-LG module, a powerful clustering tool based on Latent GOLD® 5.0:

Latent class analysis (LCA) involves the construction of latent classes (LC) which are unobserved (latent) subgroups or segments of cases. The latent classes are constructed based on the observed (manifest) responses of the cases on a set of indicator variables. Cases within the same latent class are homogeneous with respect to their responses on these indicators, while cases in different latent classes differ in their response patterns. Formally, latent classes are represented by K distinct categories of a nominal latent variable X. Since the latent variable is categorical, LC modeling differs from more traditional latent variable approaches such as factor analysis, structural equation models, and random-effects regression models since these approaches are based on continuous latent variables.

XLSTAT-LG contains separate modules for estimating two different model structures - LC Cluster models and LC Regression models - which are useful in somewhat different application areas. To better distinguish the output across modules, latent classes are labeled 'clusters' for LC Cluster models and 'classes' for LC Regression models. In this manual we also refer to latent classes using the term 'segments'.

The LC Cluster Model:

- Includes a nominal latent variable X with K categories, each category representing a cluster.
- Each cluster contains a homogeneous group of persons (cases) who share common interests, values, characteristics, and/or behavior (i.e., share common model parameters).
- These interest, values, characteristics, and/or behavior constitute the observed variables (indicators) Y upon which the latent clusters are derived.

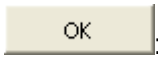
Advantages over more traditional ad-hoc types of cluster analysis methods include model selection criteria and probability-based classification. Posterior membership probabilities are estimated directly from the model parameters and used to assign cases to the modal class - the class for which the posterior probability is highest.

A special feature of LC cluster models is the ability to obtain an equation for calculating these posterior membership probabilities directly from the observed variables (indicators). This equation can be used to score new cases based on a LC cluster model estimated previously. That is, the equation can be used to classify new cases into their most likely latent class as a function of the observed variables. This feature is unique to LC models – it is not available with any other clustering technique.

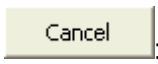
The scoring equation is obtained as a special case of the more general Step3 methodology for LC cluster models (Vermunt, 2010). In Step1, model parameter estimates are obtained. In Step2, cases are assigned to classes based on their posterior membership probabilities. In Step3, the latent classes are used as predictors or dependent variables in further analyses. For further details, see Section 2.3 (Step3 Scoring) in Vermunt and Magidson (2013b).

Dialog box

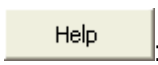
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab:

Observations/variables table:

Continuous: Select the continuous variable(s). The data must be continuous. If the 'Column labels' option is activated make sure that the headers of the variable(s) have also been selected.

Nominal: Select the nominal variable(s). The data must be nominal. If the 'Column labels' option is activated make sure that the headers of the variable(s) have also been selected.

Ordinal: Select the ordinal variable(s). The data must be numeric. If the 'Column labels' option is activated make sure that the headers of the variable(s) have also been selected.

Direct effects: Activate this option if you want to specify a direct effect to be included in the model. After specifying your model and clicking "OK" from the dialog box, an interactions box will pop up. All pairs of variables eligible for a direct effect parameter appear. To include a direct effect, click in the check-box and a check appears. Direct effect parameters will be estimated for the pairs of variables that have been so selected (direct effect check-box equals on). The inclusion of direct effects is one way to relax the assumption of local dependence.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if labels are available for the N observations. Then select the corresponding data. If the 'Column labels' option is activated you need to include a header in the selection.

With repeated measures data (multiple records per case) the Observation labels variable serves as a case ID variable, which groups the records from each case together so that they are assigned to the same fold during cross-validation. If this option is *not* activated, labels for the observations are automatically generated by XLSTAT (Obs1, Obs2 …), so that each case contains a single record.

Case weights: Activate this option if you want to weight the observations. If you do not activate this option, all weights are set to 1. The weights must be non-negative values. Setting a case weight to 2 is equivalent to repeating the same observation twice. If the 'Variable labels' option is activated, make sure that the header (first row) has also been selected.

Number of clusters:

from: Enter a number between 1-25.

to: Enter a number between 1-25.

Note: To specify a fixed number of clusters K : use from K to K. For example, to estimate a 2 class model: from 2 to 2.

Use separate sheets: Activate this option if you want the program to produce separate sheets for each cluster model estimated. A separate sheet with summary statistics for all models estimated will also be produced.

Options tab:

Parameter estimation occurs using an iterative algorithm which begins using the Expectation-Maximization (EM) algorithm until either the maximum number of EM iterations (Iterations EM) or the EM convergence criterion (Tolerance(EM)) is reached. Then, the program switches to perform Newton Raphson (NR) iterations which continue until the maximum number of NR iterations (Iterations Newton-Raphson) or the overall converge criterion (Tolerance) is reached. The program also stops iterating when the change in the log-posterior is negligible (smaller than 10^{-12}). A warning is given if one of the elements of the gradient is larger than 10^{-3}:

Sometimes, for example in the case of models with many parameters, it is more efficient to use only the EM algorithm. This is accomplished by setting Iterations Newton-Raphson to 0. With very large models, one may also consider suppressing the computation of standard errors (and associated Wald statistics) in the Output tab.

Convergence:

Tolerance(EM): Expectation-Maximization (EM) Tolerance is the sum of absolute relative changes of parameter values in a single iteration as long as the EM algorithm is used. It determines when the program switches from EM to Newton-Raphson (if the NR iteration limit has been set to > 0). Increasing the EM Tolerance will switch faster from EM to NR. To change this option, click the value to highlight it, then type in a new value. You may enter any non-negative real number. The default is 0.01. Values between 0.01 and 0.1 (1% and 10%) are reasonable.

Tolerance: Overall Tolerance (Tolerance) is the sum of absolute relative changes of parameter values in a single iteration. It determines when the program stops its iteration. The default is $1.0x10^{-8}$ which specifies a tight convergence criterion. To change this option, click the value to highlight it, then type in a new value. You may enter any non-negative real number.

Note: when only EM iterations are used, the effective tolerance is the maximum of Tolerance(EM) and Overall Tolerance.

Iterations:

EM: Maximum number of EM iterations. The default is 250. If the model does not converge after 250 iterations, this value should be increased. You also may want to increase this value if you set Newton-Raphson iterations = 0. To change this option, click the value to highlight it, then type in a new value. You may enter any non-negative integer.

Newton-Raphson: Maximum number of NR iterations. The default is 50. If the model does not converge after 50 iterations, this value should be increased. To change this option, click the value to highlight it, then type in a new value. You may enter any non-negative integer. A value of 0 is entered to direct XLSTAT-LG to use only EM, which may produce faster convergence in models with many parameters or in models that contain continuous indicators.

Start v alues:

The best way to prevent ending up with a local solution is the use of multiple sets of starting values since different sets of starting values may yield solutions with different log-posterior values. The use of such multiple sets of random starting values is automated. This procedure increases considerably the probability of finding the global solution, but in general does not guarantee that it will be found in a single run. To reduce the likelihood of obtaining a local solution, the following options can be used to either increasing the number of start sets, the number of iterations per set, or both.

Random s ets: The default is 16 for the number of random sets of starting values to be used to start the iterative estimation algorithm. Increasing the number of sets of random starting values for the model parameters reduces the likelihood of converging to a local (rather than global) solution. To change this option, click the value to highlight it, then type in a new value. You may enter any non-negative integer. Using either the value 0 or 1 results in the use of a single set of starting values.

Iterations: This option allows specification of the number of iterations to be performed per set of start values. XLSTAT-LG first performs this number of iterations within each set and subsequently twice this number within the best 10% of the start sets. For some models, many more than 50 iterations per set may need to be performed to avoid local solutions.

Seed (random numbers): The default value of 123456789 means that the Seed is obtained during estimation using a pseudo random number generator based on clock time. Specifying a non-negative integer different from 0, yields the same result each time.

To specify a particular numeric seed (such as the Best Start Seed reported in the Model Summary Output for a previously estimated model), click the value to highlight it, then type in a non-negative integer. When using the Best Start Seed, be sure to deactivate the Random Sets option (using Random Sets = 0).

Tolerance: Indicates the convergence criterion to be used when running the model of interest with the various start sets. The definition of this tolerance is the same as the one that is used for

the EM and Newton-Raphson Iterations.

Bayes Constants:

The Bayes options can be used to eliminate the possibility of obtaining boundary solutions. You may enter any non-negative real value. Separate Bayes constants can be specified for three different situations:

Latent: The default is 1. Increase the value to increase the weight allocated to the Dirichlet prior which is used to prevent the occurrence of boundary zeroes in estimating the latent distribution. The number can be interpreted as a total number of added cases that is equally distributed among the classes (and the covariate patterns). To change this option, click the value to highlight it, then type in a new value.

Categorical: The default is 1. Increase the value to increase the weight allocated to the Dirichlet prior which is used in estimating multinomial models with variables specified as Ordinal or Nominal. This number can be interpreted as a total number of added cases to the cells in the models for the indicators to prevent the occurrence of boundary solutions. To change this option, click the value to highlight it, then type in a new value.

Error v ariance: The default is 1. Increase the value to increase the weight allocated to the inverse-Wishart prior which is used in estimating the error variance-covariance matrix in models for continuous dependent variables or indicators. The number can be interpreted as the number of pseudo-cases added to the data, each pseudo-case having a squared error equal to the total variance of the indicator concerned. Such a prior prevents variances of zero from occurring. To change this option, click the value to highlight it, then type in a new value.

For technical details, see section 7.3 of Vermunt and Magidson (2013a).

Cluster Independent:

Error (Co)variances: This option indicates that the error covariances are restricted to be equal across classes (class independent). Note that this option only applies to pairs of continuous indicators for which direct effects have been included in the model (see the Direct Effects option in the General tab).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Statistics: Activate this option to display the following statistics about the model(s).

Chi-squared: Activate this option to display various chi-square based statistics related to model fit.

Log-likelihood: Activate this option to display log-likelihood statistics.

Classification: Activate this option to display the Classification Table, which cross-tabulates modal and probabilistic class assignment.

Profile: Activate this option to display the probabilities or means associated with each Indicator.

- The first row of numbers shows how large each cluster is.
- The body of the table contains (marginal) conditional probabilities that show how the clusters are related to the Nominal or Ordinal variables. These probabilities sum to 1 within each cluster (column).
- For indicators specified as Continuous, the body of the table contains means (rates) instead of probabilities. For indicators specified as Ordinal, means are displayed in addition to the conditional probabilities.

Standard Errors: Activate this option to display the standard errors (and associated Wald statistics). The standard (Hessian) computation method makes use of the second-order derivatives of the log-likelihood function called the Hessian matrix.

Bivariate Residuals: Activate this option to display the bivariate residuals for the model

Frequencies / Residuals: Activate this option to display the observed and expected frequencies along with the standardized residuals for a model. This output is not available if at least one indicators is continuous. This output is not reported in the case 1 or more continuous indicators.

Iteration Details: Activate this option to display technical information associated with the performance of the estimation algorithm, such as log- posterior and log-likelihood values at convergence:

- EM algorithm,
- Newton-Raphson algorithm.

When applicable, this file also contains warning messages concerning non- convergence, unidentified parameters and boundary solutions.

Scoring Equation: Activate this option to display the scoring equation, consisting of regression coefficients associated with the multinomial logit model. The resulting scores are predicted logits

associated with each latent class t. For example, for responses $Y_1 = j, Y_2 = k, Y_3 = m, Y_4 = s$ to 4 nominal indicators, the logit associated with cluster t is:

$$Logit(t) = a[t] + b_1[j,t] + b_2[k,t] + b_3[m,t] + b_4[s,t]$$

Thus, to obtain the posterior membership probabilities for latent class t_0 given this response pattern, use the following formula:

$$Prob(classe[t = t_0]|Y_1 = j, Y_2 = k, Y_3 = m, Y_4 = s) = exp\left(\frac{Logit[t_0]}{\sum_t exp(Logit[t]}\right)$$
$$= \frac{exp(a[t_0] + b_1[j,t_0] + b_2[k,t_0] + b_3[m}{\sum_t exp(a[t] + b_1[j,t] + b_2[k,t] + b_3[}$$

For further details, see the tutorial "Using XLSTAT-LG to estimate latent class cluster models".

Classification: Activate this option to display a table containing the posterior membership probability and the modal assignment for each of the cases based on the current model.

Charts tab:

Profile p lot: The profile plot is constructed from the conditional probabilities for the nominal variables and means for the other indicators as displayed in the columns of the Profile table. The quantities associated with the selected clusters are plotted and connected. For the scale types ordinal, continuous, count, and numeric covariate, prior to plotting the class-specific means, they are re-scaled to always lie within the 0-1 range. Scaling of these "0-1 Means" is accomplished by subtracting the lowest observed value from the class-specific means and dividing the results by the range, which is simply the difference between the highest and the lowest observed value. The advantage of such scaling is that these numbers can be depicted on the same scale as the class-specific probabilities for nominal variables. For nominal variables containing more than 2 categories, all categories are displayed simultaneously. For dichotomous variables specified as nominal, by default only the last category is displayed.

Results

Summary Sheet

Summary (descriptive) statistics: For the dependent variables and the quantitative explanatory variables, XLSTAT displays the number of observations, the number of observations with missing data, the number of observations with no missing data, the mean, and the unbiased standard deviation. For the nominal explanatory variables, the number and frequency of cases belonging to each level are displayed.

Summary Statistics:

- **Model Name**: The models are named after the number of classes the model contains.

- **LL**: The likelihood-ratio goodness-of-fit value for the current model.
- **BIC(LL), AIC(LL), AIC3(LL)**: BIC, AIC and AIC3 (based on LL). In addition to model fit, these statistics take into account the parsimony (df or Npar) of the model. When comparing models, the lower the BIC, AIC and AIC3 value the better the model.
- **Npar**: Number of parameters.
- **L²**: Likelihood ratio χ^2. Not available if the model contains 1 or more continuous indicators.
- **df**: Degrees of freedom for L^2.
- **p-value**: Model fit p-value for L^2.
- **Class.Err.**: Expected classification error. The expected proportion of cases misclassified when classification of cases is based on modal assignment (i.e., assigned to the class having the highest membership probability). The closer this value is to 0 the better.

Model Output Sheet

Model Summary Statistics:

Model:

- **Number of cases**: This is the number of cases used in model estimation. This number may be less than the original number of cases on the data file if missing cases have been excluded.
- **Number of replications**: Total number of observations
- **Number of parameters (Npar)**: This is the number of distinct parameters estimated.
- **Seed (random numbers)**: The seed required to reproduce this model.
- **Best seed**: The single best seed that can reproduce this model more quickly using the number of starting sets =0.

Estimation summary:

- **EM iterations**: number of EM iterations used.
- **Log-posterior**: Log-posterior value.
- **L²**: The likelihood-ratio goodness-of-fit value for the current model.
- **Final convergence value**: Final convergence value.
- **Newton-Raphson iteration**: Number of Newton-Raphson iterations used.

- **Log-posterior**: Log-posterior value.
- **L²**: The likelihood-ratio goodness-of-fit value for the current model.
- **Final convergence value**: Final convergence value.

Chi-Square statistics:

- **Degrees of freedom (df)**: The degrees of freedom for the current model.
- **L²**: The likelihood-ratio goodness-of-fit value for the current model. If the bootstrap p-value for the L^2 statistic has been requested, the results will be displayed here.
- **X² and Cressie-Read**: These are alternatives to L^2 that should yield a similar p-value according to large sample theory if the model specified is valid and the data is not sparse.
- **BIC, AIC, AIC3 and CAIC (based on L²)**: In addition to model fit, these statistics take into account the parsimony (df or Npar) of the model. When comparing models, the lower the BIC, AIC, AIC3 and CAIC value the better the model.
- **SABIC (based on L²)**: Sample size adjusted BIC, an information criterion similar to BIC but with $log(N)$ replaced by $log(\frac{(N+2)}{24})$.
- **Dissimilarity Index**: A descriptive measure indicating how much the observed and estimated cell frequencies differ from one another. It indicates the proportion of the sample that needs to be movedto another cell to get a perfect fit.

Log-likelihood statistics:

- **Log-likelihood (LL)**: LN(Likelihood) displayed here.
- **Log-prior:** this is the term in the function maximized in the parameter estimation that is associated with the Bayes constants. This term equals 0 if all Bayes constants are set to 0.
- **Log-posterior:** this is the term in the function that is maximized in the parameter estimation. The value of the log-posterior function is obtained as the sum of the log-likelihood and log-prior values.
- **BIC, AIC, AIC3 and CAIC (based on LL)**: these statistics (information criteria) weight fit and parsimony by adjusting the LL to account for the number of parameters in the model. The lower the value, the better the model.
- **SABIC (based on LL)**: Sample size adjusted BIC, an information criterion similar to BIC but with $log(N)$ replaced by $log(\frac{(N+2)}{24})$.

Classification statistics:

- **Classification errors**: When classification of cases is based on modal assignment (to the class having the highest membership probability), the proportion of cases that are estimated to be misclassified is reported by this statistic. The closer this value is to 0 the better.
- **Reduction of errors (Lambda), Entropy R², Standard R²**: These pseudo R^2 statistics indicate how well one can predict class memberships based on the observed variables (indicators and covariates). The closer these values are to 1 the better the predictions.
- **Classification Log-likelihood**: Log-likelihood value under the assumption that the true class membership is known.
- **EN**: Entropy.
- **CLC**: CL*2
- **AWE**: Similar to BIC, but also takes classification performance into account.
- **ICL-BIC**: BIC-2*En

Classification table:

- **Modal table**: Cross-tabulates modal class assignments.
- **Proportional table**: Cross-tabulates probabilistic class assignments.

Profile:

- **Cluster Size**: Size of each cluster
- **Indicators**: The body of the table contains (marginal) conditional probabilities that show how the clusters are related to the Nominal or Ordinal indicator variables. These probabilities sum to 1. For indicators specified as Continuous, the body of the table contains means instead of probabilities. For indicators specified as Ordinal, means are displayed in addition to the conditional probabilities within each cluster (column).
- **s.e. (standard errors)**: standard errors for the (marginal) conditional probabilities.
- **Profile plot**: The probabilities and means that appear in the Profile Output, are displayed graphically in the Profile Plot

Frequencies / Residuals:

Table of observed vs. estimated expected frequencies (and residuals). Note: Residuals having magnitude greater than 2 are statistically significant. This output is not reported in the case of 1 or more continuous indicators.

Bivariate residuals:

- **Indicators**: a table containing the bivariate residuals (BVRs) for a model. Large BVRs suggest violation of the local independence assumption.

Scoring equation: regression coefficients associated with the multinomial logit model.

Classification: Outputs for each observation the posterior class memberships and the modal assignment based on the current model.

Estimation Warnings

WARNING: negative number of degrees of freedom.

This warning indicates that the model contains more parameters than cell counts. A necessary (but not sufficient) condition for identification of the parameters of a latent class model is that the number of degrees of freedom is nonnegative. This warning thus indicates that the model is not identified. The remedy is to use a model with fewer latent classes.

WARNING: # boundary or non-identified parameter(s)

This warning is derived from the rank of the information matrix (Hessian or its outer-product approximation). When there are non-identified parameters, the information matrix will not be full rank. The number reported is the rank deficiency, which gives an indication of the number of non-identified parameters.

Note that there are two problems associated with this identification check. The first is that boundary estimates also yield rank deficiencies. In other words, when there is a rank deficiency, we do not know whether it is caused by boundaries or non-identified parameters. The XLSTAT-LG Bayes Constants prevent boundaries from occurring, which solves the first problem related to this message. However, a second problem is that this identification check cannot always detect non-identification when Bayes Constants are used; that is, Bayes Constants can make an otherwise non-identified model appear to be identified.

WARNING: maximum number of iterations reached without convergence

This warning is provided if the maximum specified EM and Newton-Raphson iterations are reached without meeting the tolerance criterion. If the (by default very strict) tolerance is almost reached, the solution is probably be ok. Otherwise, the remedy is to reestimate the model with a sharper EM tolerance and/or more EM iterations, which makes sure that the switch from EM to Newton-Raphson occurs later. The default number of 50 Newton-Raphson iterations will generally be more than sufficient.

Example

A tutorial on how to use latent class clustering is available on XLSTAT Help Center:

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

References

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

Latent class regression

This tool is part of the XLSTAT-LG module. Use this tool to classify cases into meaningful clusters (latent classes) that differ on one or more parameters from latent class (LC) Cluster models. LC Cluster models classify based on combinations of continuous and/or categorical (nominal or ordinal) variables.

In this section:

Description

The latent class regression feature of XLSTAT is part of the XLSTAT-LG module, a powerful clustering tool based on Latent GOLD® 5.0:

Latent class analysis (LCA) involves the construction of latent classes (LC) which are unobserved (latent) subgroups or segments of cases. The latent classes are constructed based on the observed (manifest) responses of the cases on a set of indicator variables. Cases within the same latent class are homogeneous with respect to their responses on these indicators, while cases in different latent classes differ in their response patterns. Formally, latent classes are represented by K distinct categories of a nominal latent variable X. Since the latent variable is categorical, LC modeling differs from more traditional latent variable approaches such as factor analysis, structural equation models, and random-effects regression models since these approaches are based on continuous latent variables.

XLSTAT-LG contains separate modules for estimating two different model structures - LC Cluster models and LC Regression models - which are useful in somewhat different application areas. To better distinguish the output across modules, latent classes are labeled 'clusters' for LC Cluster models and 'classes' for LC Regression models. In this manual we also refer to latent classes using the term 'segments'.

The LC Regression Model:

- Is used to predict a dependent variable as a function of predictor variables (Regression model).
- Includes a K-category latent variable X to cluster cases (LC model)

- Each category represents a homogeneous subpopulation (segment) having identical regression coefficients (LC Regression Model).
- Each case may contain multiple records (Regression with repeated measurements).
- The appropriate model is estimated according to the scale type of the dependent variable:
- Continuous - Linear regression model (with normally distributed residuals)
- Nominal (with more than 2 levels) - Multinomial logistic regression
- Ordinal (with more than 2 ordered levels) - Adjacent-category ordinal logistic regression model
- Count: Log-linear Poisson regression
- Binomial Count: Binomial logistic regression model

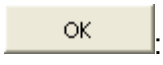
Note that a dichotomous dependent variable can be analyzed using either nominal, ordinal, or a binomial count as its scale type without any difference in the model results.

For either of the two model structures:

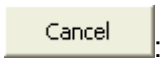
- Diagnostic statistics are available to help determine the number of latent classes, clusters, or segments
- For models containing $K > 1$ classes, covariates can be included in the model to improve classification of each case into the most likely segment.

Dialog box

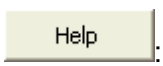
The dialog box is divided into several tabs that correspond to a variety of options ranging from the selection of data to the display of results. You will find below the description of the various elements of the dialog box.



: Click this button to start the computations.



: Click this button to close the dialog box without doing any computation.



: Click this button to display the help.


: Click this button to reload the default options.

 : Click this button to delete the data selections.

 : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

General tab :

Y / Dependent variables : Select the dependent variable here. If the 'Column labels' option is activated make sure that the headers of the variable(s) have also been selected.

Note: In case multiple dependent variables are selected, multiple independent regression analysis are performed. Separate output for each dependent variable will be provided and only a single scale type can be selected for all the dependent variables.

Response type: Select the scale type of the dependent variable. The dependent variable may be Nominal, Ordinal, Continuous, Binomial, or Count.

- **Nominal.** This setting should be used for categorical variables where the categories have no natural ordering. If the dependent variable is set to Nominal, the multinomial logit model is used.
- **Ordinal.** This setting should be used for categorical variables where the categories are ordered (either from high to low or low to high). The adjacent category logit model, also known as the baseline category logit model is specified.
- **Continuous.** This setting should be used when the variable is continuous. If the dependent variable is set to Continuous, the normal linear Regression model is used.
- **Binomial.** This setting should be used when the variable represents binomial counts. If the dependent variable is set to Binomial Count, the binomial model is used and you can also specify a variable to be used as an exposure (see Exposure). During the scan, the program checks to make sure that the exposure, if specified, is larger than any observed count.
- **Counts.** This setting should be used when the variable represents Poisson counts. If the dependent variable is set to Count, the Poisson model is used and you can also specify an additional variable to be used as an exposure (see Exposure).

Exposure. The Exposure field is active only if the scale type for the dependent variable has been specified to be Binomial or Count. (For other scale types, no exposure variable is used.)

For dependent variables specified as Binomial or Count, the exposure is specified by designating a variable as the exposure variable or, if no such variable is designated, by entering a value in the exposure constant box which appears to the right of the Exposure variable box. The use of an exposure variable allows the exposure to vary over cases.

By default, the value in the Exposure constant box is 1, a value often used to represent the Poisson exposure. To change the exposure constant, highlight the value in the exposure constant box and type in the desired value. Alternatively, you can select an exposure variable

When the scale type is specified as Binomial, the value of the dependent variable represents the number of 'successes' in N trials. In this case, the exposure represents the number of trials (the values for N), and hence should never take on a value lower than the value of the dependent variable and hence typically should be higher than the default constant of 1. Before the actual model estimation, XLSTAT-LG checks each case and will provide a warning message if this condition is not met for one or more cases. An exposure variable should be designated if the number of trials is not the same for all cases.

Explanatory variables. Select any variable(s) to be used as predictors of the dependent variable. Predictors may be treated as Nominal or Numeric. If no predictors are selected, the model will contain an intercept only.

- **Numeric.** This setting should be used for an ordinal or continuous covariate or predictor.
- **Nominal.** This setting should be used for categorical variables where the categories have no natural ordering.

Range: Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

Sheet: Activate this option to display the results in a new worksheet of the active workbook.

Workbook: Activate this option to display the results in a new workbook.

Column labels: Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

Observation labels: Activate this option if labels are available for the N observations. Then select the corresponding data. If the ‘Column labels’ option is activated you need to include a header in the selection.

With repeated measures data (multiple records per case) the Observation labels variable serves as a case ID variable, which groups the records from each case together so that they are assigned to the same fold during cross-validation. If this option is *not* activated, labels for the observations are automatically generated by XLSTAT (Obs1, Obs2 …), so that each case contains a single record.

Replication weights: Activate this option to assign a Replication Weight. A common application of replication weights is in the estimation of certain kinds of allocation models, where respondents assign a fixed number of points to each of J alternatives. For each case, the assigned points are used as replication weights to weight each of J responses. A weighted multinomial logit model is estimated.

Case weights: Activate this option if you want to weight the observations. If you do not activate this option, all weights are set to 1. The weights must be non-negative values. Setting a case weight to 2 is equivalent to repeating the same observation twice. If the 'Variable labels' option is activated, make sure that the header (first row) has also been selected.

Number of clusters:

from: Enter a number between 1-25.

to: Enter a number between 1-25.

Note: To specify a fixed number of clusters K : use from K to K. For example, to estimate a 2 class model: from 2 to 2.

Use separate sheets: Activate this option if you want the program to produce separate sheets for each cluster model estimated. A separate sheet with summary statistics for all models estimated will also be produced.

Options tab:

Parameter estimation occurs using an iterative algorithm which begins using the Expectation-Maximization (EM) algorithm until either the maximum number of EM iterations (Iterations **EM**) or the EM convergence criterion (**Tolerance(EM)**) is reached. Then, the program switches to perform Newton Raphson (NR) iterations which continue until the maximum number of NR iterations (Iterations **Newton-Raphson**) or the overall converge criterion (**Tolerance**) is reached. The program also stops iterating when the change in the log-posterior is negligible (smaller than 10^{-12}). A warning is given if one of the elements of the gradient is larger than 10^{-3}:

Sometimes, for example in the case of models with many parameters, it is more efficient to use only the EM algorithm. This is accomplished by setting Iterations Newton-Raphson to 0. With very large models, one may also consider suppressing the computation of standard errors (and associated Wald statistics).

Convergence:

Tolerance(EM): Expectation-Maximization (EM) Tolerance is the sum of absolute relative changes of parameter values in a single iteration as long as the EM algorithm is used. It determines when the program switches from EM to Newton-Raphson (if the NR iteration limit has been set to > 0). Increasing the EM Tolerance will switch faster from EM to NR. To change this option, click the value to highlight it, then type in a new value. You may enter any non-negative real number. The default is 0.01. Values between 0.01 and 0.1 (1% and 10%) are reasonable.

Tolerance: Overall Tolerance (Tolerance) is the sum of absolute relative changes of parameter values in a single iteration. It determines when the program stops its iteration. The default is $1.0x10^{-8}$ which specifies a tight convergence criterion. To change this option, click the value to highlight it, then type in a new value. You may enter any non-negative real number.

Iterations:

EM: Maximum number of EM iterations. The default is 250. If the model does not converge after 250 iterations, this value should be increased. You also may want to increase this value if you set Newton-Raphson iterations = 0. To change this option, click the value to highlight it, then type in a new value. You may enter any non-negative integer.

Newton-Raphson: Maximum number of NR iterations. The default is 50. If the model does not converge after 50 iterations, this value should be increased. To change this option, click the value to highlight it, then type in a new value. You may enter any non-negative integer. A value of 0 is entered to direct XLSTAT-LG to use only EM, which may produce faster convergence in models with many parameters or in models that contain continuous indicators.

Start values:

The best way to prevent ending up with a local solution is the use of multiple sets of starting values since different sets of starting values may yield solutions with different log-posterior values. The use of such multiple sets of random starting values is automated. This procedure increases considerably the probability of finding the global solution, but in general does not guarantee that it will be found in a single run. To reduce the likelihood of obtaining a local solution, the following options can be used to either increasing the number of start sets, the number of iterations per set, or both.

Random sets: The default is 16 for the number of random sets of starting values to be used to start the iterative estimation algorithm. Increasing the number of sets of random starting values for the model parameters reduces the likelihood of converging to a local (rather than global) solution. To change this option, click the value to highlight it, then type in a new value. You may enter any non-negative integer. Using either the value 0 or 1 results in the use of a single set of starting values.

Iterations: This option allows specification of the number of iterations to be performed per set of start values. XLSTAT-LG first performs this number of iterations within each set and subsequently twice this number within the best 10% of the start sets. For some models, many more than 50 iterations per set may need to be performed to avoid local solutions.

Seed (random numbers): The default value of 123456789 means that the Seed is obtained during estimation using a pseudo random number generator based on clock time. Specifying a non-negative integer different from 0, yields the same result each time.

To specify a particular numeric seed (such as the Best Start Seed reported in the Model Summary Output for a previously estimated model), click the value to highlight it, then type in a non-negative integer. When using the Best Start Seed, be sure to deactivate the Random Sets option (using Random Sets = 0).

Tolerance: Indicates the convergence criterion to be used when running the model of interest with the various start sets. The definition of this tolerance is the same as the one that is used for the EM and Newton-Raphson Iterations.

Bayes Constants:

The Bayes options can be used to eliminate the possibility of obtaining boundary solutions. You may enter any non-negative real value. Separate Bayes constants can be specified for three different situations:

Latent: The default is 1. Increase the value to increase the weight allocated to the Dirichlet prior which is used to prevent the occurrence of boundary zeroes in estimating the latent distribution. The number can be interpreted as a total number of added cases that is equally distributed among the classes (and the covariate patterns). To change this option, click the value to highlight it, then type in a new value.

Categorical: The default is 1. Increase the value to increase the weight allocated to the Dirichlet prior which is used in estimating multinomial models with variables specified as Ordinal or Nominal. This number can be interpreted as a total number of added cases to the cells in the models for the indicators to prevent the occurrence of boundary solutions. To change this option, click the value to highlight it, then type in a new value.

Error v ariance: The default is 1. Increase the value to increase the weight allocated to the inverse-Wishart prior which is used in estimating the error variance-covariance matrix in models for continuous dependent variables or indicators. The number can be interpreted as the number of pseudo-cases added to the data, each pseudo-case having a squared error equal to the total variance of the indicator concerned. Such a prior prevents variances of zero from occurring. To change this option, click the value to highlight it, then type in a new value.

For technical details, see section 7.3 of Vermunt and Magidson (2013a).

Class Independent:

Various restrictions are available for intercepts and predictor effects. In addition, for models with continuous dependent variables, restrictions are available for error variances.

- **Error variances:** This option indicates that the error covariances are restricted to be equal across classes (class independent).
- **Predictors (1 or more).** This option indicates that the predictors are restricted to be equal across classes (class independent).
- **Intercept.** This option indicates that the intercept is restricted to be equal across classes (class independent).

Missing data tab:

Do not accept missing data: Activate this option so that XLSTAT prevents the computations from continuing if missing data have been detected.

Remove observations: Activate this option to remove the observations with missing data.

Outputs tab:

Descriptive statistics: Activate this option to display descriptive statistics for the variables selected.

Statistics: Activate this option to display the following statistics about the model(s).

Chi-square: Activate this option to display various chi-square based statistics related to model fit.

Log-likelihood: Activate this option to display log-likelihood statistics.

Classification: Activate this option to display the Classification Table, which cross-tabulates modal and probabilistic class assignment.

Parameters:

Standard errors: Activate this option to display the standard errors of the parameters. The standard (Hessian) computation method makes use of the second-order derivatives of the log-likelihood function called the Hessian matrix.

Wald t ests: Activate this option to display the Wald statistics.

Frequencies / Residuals: Activate this option to display the observed and expected frequencies along with the standardized residuals for a model. This output is not available if at least one indicators is continuous. This output is not reported in the case 1 or more continuous indicators.

Iteration details: Activate this option to display technical information associated with the performance of the estimation algorithm, such as log- posterior and log-likelihood values at convergence:

- EM algorithm,
- Newton algorithm.

When applicable, this file also contains warning messages concerning non- convergence, unidentified parameters and boundary solutions.

Estimated values: Activate this option to display the predicted values information (the probability of responding to each category) to the data. The following variables (and variable names) will be shown:

- pred_1 - the predicted prob of responding in the first category
- pred_2 - the predicted prob of responding in the second category
- pred_dep - the predicted value (weighted average of the category scores, with the predicted probs as the weights)

Classification: Activate this option to display a table containing the posterior membership probability and the modal assignment for each of the cases based on the current model.

Nominal coding:

Effect (default). By default, the Parameter Output contains effect coding for nominal indicators, dependent variable, active covariates and the latent classes (clusters).

Use either of these options to change to dummy coding.

a1=0 (Dummy First). Selection of this option causes dummy coding to be used with the first category serving as the reference category.

an=0 (Dummy Last). Selection of this option causes dummy coding to be used with the last category serving as the reference category.

Charts tab:

Profile p lot: Activate this option to display the profile plot.

Results

Summary Sheet

Summary (descriptive) statistics: For the dependent variables and the quantitative explanatory variables, XLSTAT displays the number of observations, the number of observations with missing data, the number of observations with no missing data, the mean, and the unbiased standard deviation. For the nominal explanatory variables, the number and frequency of cases belonging to each level are displayed.

Summary Statistics:

- **Model Name:** The models are named after the number of classes the model contains.
- **LL:** The likelihood-ratio goodness-of-fit value for the current model.
- **BIC(LL), AIC(LL), AIC3(LL):** BIC, AIC and AIC3 (based on LL). In addition to model fit, these statistics take into account the parsimony (df or Npar) of the model. When comparing models, the lower the BIC, AIC and AIC3 value the better the model.
- **Npar:** Number of parameters.
- **L²:** Likelihood ratio chi-squared. Not available if the model contains 1 or more continuous indicators.
- **df:** Degrees of freedom for L^2.
- **p-value:** Model fit p-value for L^2.

- **Class.Err.:** Expected classification error. The expected proportion of cases misclassified when classification of cases is based on modal assignment (i.e., assigned to the class having the highest membership probability). The closer this value is to 0 the better.

Model Output Sheet

Model Summary Statistics:

Model:

- **Number of cases:** This is the number of cases used in model estimation. This number may be less than the original number of cases on the data file if missing cases have been excluded.
- **Number of replications:** Total number of observations
- **Number of parameters (Npar):** This is the number of distinct parameters estimated.
- **Seed (random numbers) :** The seed required to reproduce this model.
- **Best seed:** The single best seed that can reproduce this model more quickly using the number of starting sets =0.

Estimation summary:

- **EM iterations**: number of EM iterations used.
- **Log-posterior**: Log-posterior value.
- **L^2**: The likelihood-ratio goodness-of-fit value for the current model.
- **Final convergence value**: Final convergence value.
- **Newton-Raphson iteration**: number of Newton-Raphson iterations used.
- **Log-posterior**: Log-posterior value.
- **L^2**: The likelihood-ratio goodness-of-fit value for the current model.
- **Final convergence value**: Final convergence value.

Chi-Square statistics:

- **Degrees of freedom (df)**: The degrees of freedom for the current model.
- **L^2**: The likelihood-ratio goodness-of-fit value for the current model. If the bootstrap p-value for the L2 statistic has been requested, the results will be displayed here.

- **X² and Cressie-Read**: These are alternatives to L^2 that should yield a similar p-value according to large sample theory if the model specified is valid and the data is not sparse.
- **BIC, AIC, AIC3 and CAIC (based on L²)**: In addition to model fit, these statistics take into account the parsimony (df or Npar) of the model. When comparing models, the lower the BIC, AIC, AIC3 and CAIC value the better the model.
- **SABIC (based on L²)**: Sample size adjusted BIC, an information criterion similar to BIC but with $log(N)$ replaced by $log(\frac{(N+2)}{24})$.
- **Dissimilarity Index**: A descriptive measure indicating how much the observed and estimated cell frequencies differ from one another. It indicates the proportion of the sample that needs to be movedto another cell to get a perfect fit.

Log-likelihood statistics:

Log-likelihood (LL): displayed here.

- **Log-prior**: this is the term in the function maximized in the parameter estimation that is associated with the Bayes constants. This term equals 0 if all Bayes constants are set to 0.
- **Log-posterior**: this is the function that is maximized in the parameter estimation. The value of the log-posterior function is obtained as the sum of the log-likelihood and log-prior values.
- **BIC, AIC, AIC3 and CAIC (based on LL)**: these statistics (information criteria) weight fit and parsimony by adjusting the LL to account for the number of parameters in the model. The lower the value, the better the model.
- **SABIC (based on LL)**: Sample size adjusted BIC, an information criterion similar to BIC but with $log(N)$ replaced by $log(\frac{(N+2)}{24})$.

Classification statistics:

- **Classification errors**: When classification of cases is based on modal assignment (to the class having the highest membership probability), the proportion of cases that are estimated to be misclassified is reported by this statistic. The closer this value is to 0 the better.
- **Reduction of errors (Lambda), Entropy R², Standard R²**: These pseudo R^2 statistics indicate how well one can predict class memberships based on the observed variables (indicators and covariates). The closer these values are to 1 the better the predictions.
- **Classification Log-likelihood**: Log-likelihood value under the assumption that the true class membership is known.
- **EN**: Entropy.

- **CLC**: CL*2
- **AWE**: Similar to BIC, but also takes classification performance into account.
- **ICL-BIC**: BIC-2*En

Classification table:

- **Modal table**: Cross-tabulates modal class assignments.
- **Proportional table**: Cross-tabulates probabilistic class assignments.

Prediction s tatistics:

The columns in this table correspond to:

- **Baseline**: prediction error of the baseline model (also referred to as null-model)
- **Model**: the prediction error of the estimated model.
- **R²**: the proportional reduction of errors in the estimated model compared to the baseline model

The rows in this table correspond to:

- **Squared Error**: Average prediction error based on squared error.
- **Minus Log-likelihood**: Average prediction error based on minus the log-likelihood.
- **Absolute Error**: Average prediction error based on absolute error.
- **Prediction error**: Average prediction error based on proportion of prediction errors (for categorical variables only).

For technical information, see section 8.1.5 of Vermunt and Magidson (2013a).

Prediction table: For nominal and ordinal dependent variables, a prediction table that cross-classifies observed and against estimated values is also provided.

Parameters:

- **R²**: class-specific and overall R^2 values. The overall R^2 indicates how well the dependent variable is overall predicted by the model (same measure as appearing in Prediction Statistics). For ordinal, continuous, and (binomial) counts, these are standard R^2 measures. For nominal dependent variables, these can be seen as weighted averages of separate R^2 measures for each category, where each category is represented by a dummy variable = 1 for that category and 0 for all other categories.

- **Intercept**: intercept of the linear regression equation.
- **s.e.**: standard errors of the parameters.
- **z-value**: z-test statistics corresponding to the parameter tests.
- **Wald**: Wald statistics are provided in the output to assess the statistical significance of the set of parameter estimates associated with a given variable. Specifically, for each variable, the Wald statistic tests the restriction that each of the parameter estimates in that set equals zero (for variables specified as Nominal, the set includes parameters for each category of the variable). For Regression models, by default, two Wald statistics (Wald, Wald(=)) are provided in the table when more than 1 class has been estimated. For each set of parameter estimates, the Wald(=) statistic considers the subset associated with each class and tests the restriction that each parameter in that subset equals the corresponding parameter in the subsets associated with each of the other classes. That is, the Wald(=) statistic tests the equality of each set of regression effects across classes.
- **p-value**: measures of significance for the estimates.
- **Mean**: means for the regression coefficients.
- **Std.Dev**: standard deviations for the regression coefficients.

Classification: Outputs for each observation the posterior class memberships and the modal assignment based on the current model.

Estimation Warnings

WARNING: negative number of degrees of freedom.

This warning indicates that the model contains more parameters than cell counts. A necessary (but not sufficient) condition for identification of the parameters of a latent class model is that the number of degrees of freedom is nonnegative. This warning thus indicates that the model is not identified. The remedy is to use a model with fewer latent classes.

WARNING: # boundary or non-identified parameter(s)

This warning is derived from the rank of the information matrix (Hessian or its outer-product approximation). When there are non-identified parameters, the information matrix will not be full rank. The number reported is the rank deficiency, which gives an indication of the number of non-identified parameters.

Note that there are two problems associated with this identification check. The first is that boundary estimates also yield rank deficiencies. In other words, when there is a rank deficiency, we do not know whether it is caused by boundaries or non-identified parameters. The XLSTAT-LG Bayes Constants prevent boundaries from occurring, which solves the first problem related to this message. However, a second problem is that this identification check cannot always detect non-identification when Bayes Constants are used; that is, Bayes Constants can make an otherwise non-identified model appear to be identified.

WARNING: maximum number of iterations reached without convergence

This warning is provided if the maximum specified EM and Newton-Raphson iterations are reached without meeting the tolerance criterion. If the (by default very strict) tolerance is almost reached, the solution is probably be ok. Otherwise, the remedy is to reestimate the model with a sharper EM tolerance and/or more EM iterations, which makes sure that the switch from EM to Newton-Raphson occurs later. The default number of 50 Newton-Raphson iterations will generally be more than sufficient.

WARNING: estimation procedure did not converge (# gradients larger than $1.0e-3$)

This message may be related to the previous message, in which case the same remedy may be used. If the previous message is not reported, this indicates that there is a more serious non-convergence problem. The algorithms may have gotten trapped in a very flat region of the parameters space (a saddle point). The best remedy is to re-estimate the model with a different seed, and possibly with a larger number of Start Sets and more Iterations per set.

Example

A tutorial on how to use latent class clustering is available on XLSTAT Help Center:

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

References

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