Generating box plots with XLSTAT

demoBP.xls

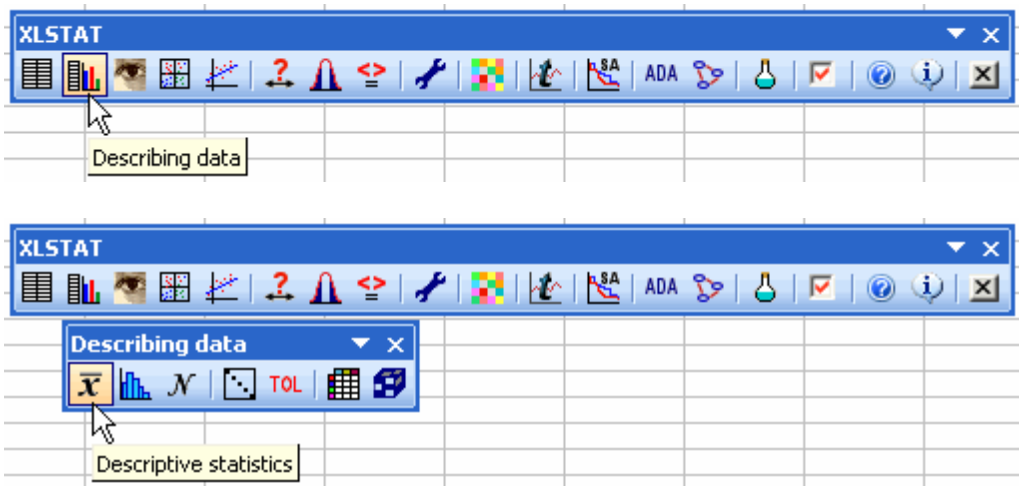
Dataset for generating a box plot

An Excel sheet with both the data and the results can be downloaded by clicking here.

The data correspond to a sample of 150 irises for which 4 variables were measured. The flowers belong to 3 different species. Fisher used this dataset, now famous, when he developed his discriminant analysis theory. In this particular example, we decided to analyze the variable "Sepal length" of the flowers and check if there are "visually" significant differences between the three species.

Setting up the dialog box for the box plot

Once XLSTAT is open, select the **XLSTAT / Describing data / Descriptive Statistics** command, or click on the corresponding button of the **Describing data** toolbar (see below).



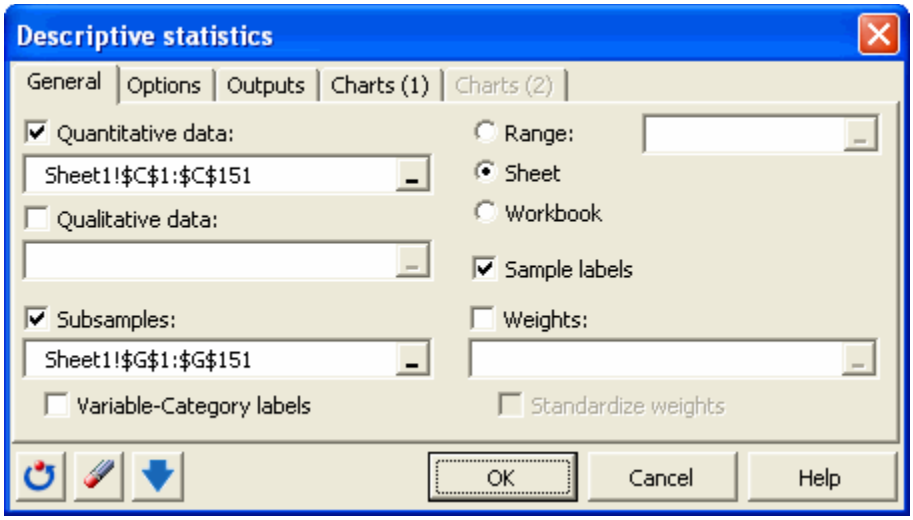
Once you have clicked on the button, the **Descriptive Statistics** dialog box appears.

The data corresponding to the variable "Sepal length" were selected on the Excel sheet. Note that for a box plot, the data must be numerical data.

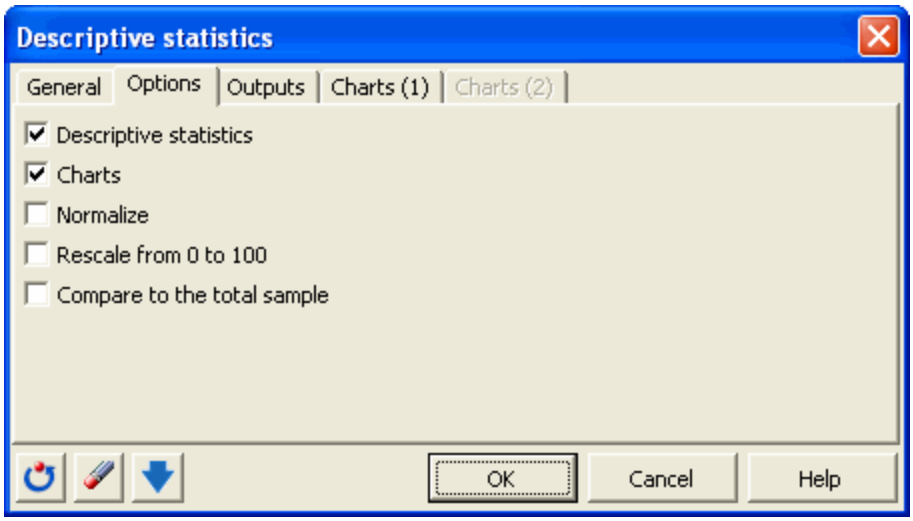
As the name of the variable was included in the selection, the **Labels included** option was also selected.

The "Species" data were selected as **sub-sample descriptor** to enable the comparison between the species.

The **Sheet** option was selected because we wanted the results displayed on a new sheet of the workbook.



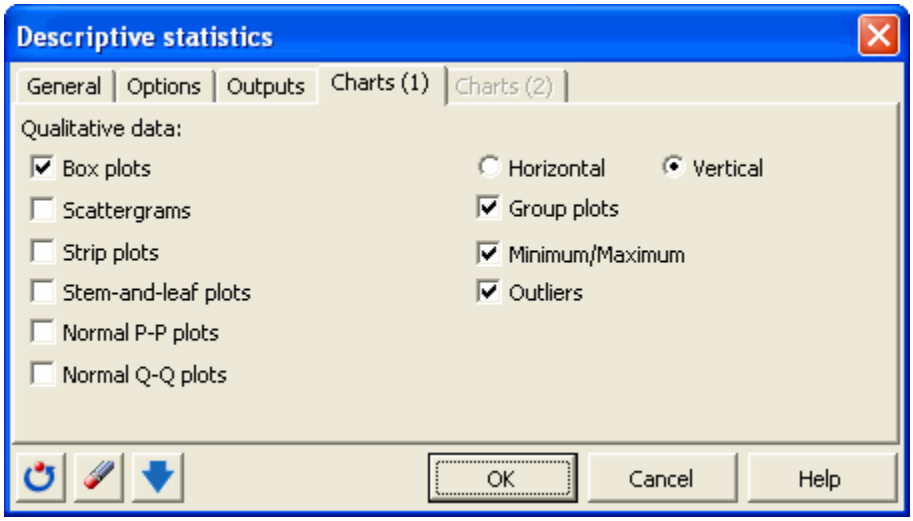
In the **Options** tab, the following options have been activated.



The **Normalize** or **Rescale** options can be used when you want to compare several variables spread over different scales - there is no need to use these in this case as we are dealing with only one variable.

In the **Charts** tab the box plots option is checked. The **Group plots** option has been chosen so that the box plots are displayed on the same chart, and not separately.

The **Minimum/Maximum** and **Outliers** options have been checked so that the corresponding values are displayed on the box plots.



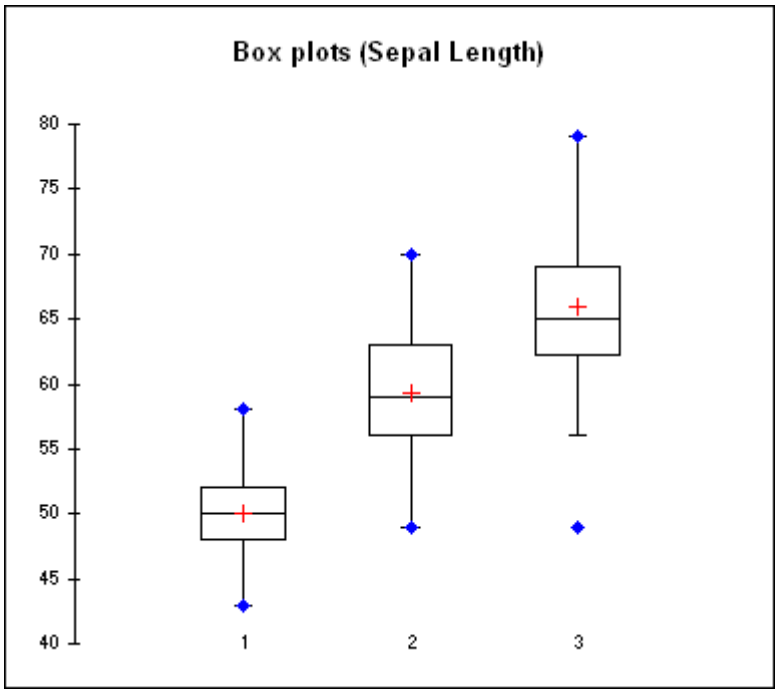
Interpreting a box plot

The results are displayed on the new sheet named "Desc". They include a full set of descriptive statistics.

Descriptive statistics (Quantitative data):

Statistic	1	2	3
No. of observations	50	50	50
No. of missing values	0	0	0
Sum of weights	50	50	50
Minimum	43.000	49.000	49.000
Maximum	58.000	70.000	79.000
Freq. of minimum	1	1	1
Freq. of maximum	1	1	1
Range	15.000	21.000	30.000
1st Quartile	48.000	56.000	62.250
Median	50.000	59.000	65.000
3rd Quartile	52.000	63.000	69.000
Sum	2503.000	2968.000	3294.000
Mean	50.060	59.360	65.880
Variance (n)	12.176	26.110	39.626
Variance (n-1)	12.425	26.643	40.434
Standard deviation (n)	3.489	5.110	6.295
Standard deviation (n-1)	3.525	5.162	6.359
Variation coefficient	0.070	0.086	0.096
Skewness (Pearson)	0.116	0.102	0.114
Skewness (Fisher)	0.120	0.105	0.118
Skewness (Bowley)	0.000	0.143	0.185
Kurtosis (Pearson)	-0.346	-0.599	-0.088
Kurtosis (Fisher)	-0.253	-0.533	0.033
Standard error of the mean	0.493	0.723	0.890
Lower bound on mean (95%)	49.068	57.908	64.091
Upper bound on mean (95%)	51.052	60.812	67.669
Mean absolute deviation	2.707	4.214	5.026
Median absolute deviation	2.000	3.500	4.000
Geometric mean	49.938	59.140	65.578
Geometric standard deviation	1.073	1.091	1.102
Harmonic mean	49.817	58.919	65.274

Then, the box plots are displayed.



It appears clearly that the Sepal length variable is different from for the three species. The blue rhombuses correspond to the minimum and maximum values.

Watch this video to see how to generate this boxplot.

http://www.youtube.com/watch?feature=player_embedded&v=MAVsGvKPZlM