# *Data handling – Outliers*

Suppose we have the file *Dataset1.xlsx* containing 3 variables:

1) *X (continuous numerical variable)*
2) *Y (discrete numerical variable)*
3) *Gender (categorical variable)*

To draw the box plot of the variable *X* or *Y*, select the data, then
INSERT ➔ GRAPHS ➔ ALL GRAPHS ➔ BOX AND WHISKER
INSERISCI ➔ GRAFICI ➔ TUTTI I GRAFICI ➔ SCATOLA E BAFFI

To delete the row containing the outlier, select the row and delete it.
To replace the outlier, see next Section.

# *Data handling – Missing values*

***Missing value of a continuous numerical variable***

Suppose we have the file *Dataset2a.xlsx* containing 3 variables:

1) *X (continuous numerical variable)*
2) *Y (discrete numerical variable)*
3) *Gender (categorical variable)*

The variable *X* has a missing value. We can use the variable *Y* and/or *Gender* to estimate the missing value.

In correspondence with the missing value of *X*, we have *Y = 7* and *Gender = F*. Therefore, select the cases (observations) with *Y = 7* and/or *Gender = F* using the filter and compute the mean (or median) of the variable *X* among the selected cases.

***Missing value of a discrete numerical variable***

Suppose we have the file *Dataset2b.xlsx* containing 3 variables:

1) *X (continuous numerical variable)*
2) *Y (discrete numerical variable)*
3) *Gender (categorical variable)*

The variable *Y* has a missing value. We can use the variable *Gender* to estimate the missing value.

In correspondence with the missing value of *Y*, we have *Gender = F*. Therefore, select the cases (observations) with *Gender = F* using the filter and identify the mode of the variable *Y* among the selected cases.

***Missing value of a categorical variable***

Suppose we have the file *Dataset2c.xlsx* containing 3 variables:

1) *X (continuous numerical variable)*
2) *Y (discrete numerical variable)*
3) *Gender (categorical variable)*

The variable *Gender* has a missing value. We can use the variable *Y* to estimate the missing value.

In correspondence with the missing value of *Gender*, we have *Y = 5*. Therefore, select the cases (observations) with *Y = 5* using the filter and identify the mode of the variable *Gender* among the selected cases.

Suppose we have the file *Dataset2d.xlsx* containing 3 variables for 235 families:

1) *Contract (categorical variable with categories C and F)*
2) *Components (discrete numerical variable)*
3) *Income (continuous numerical variable)*

The variable *Contract* has a missing value. We can use the *k-NN* technique to estimate the missing value.
Compute the distance of each family from the family with the missing value.
For instance, the distance between families A and B is given by the formula

$$d_{AB} = \sqrt{(Components_A - Components_B)^2 + (Income_A - Income_B)^2}$$

after standardizing the variables using the function **STANDARDIZE (NORMALIZZA)**.
Then, sort the observations in ascending order, according to the distance.
Select a number of families equal to $k = \sqrt{235}$.
Identify the mode of the categorical variables *Contract* in this subset of families.

# *Data handling – Inaccuracies*

Suppose we have the file *Dataset3.xlsx* containing 3 variables:

1) *X (continuous numerical variable)*
2) *Y (discrete numerical variable)*
3) *Gender (categorical variable)*

To check the number of categories of the variable *Gender*, use the table with the frequencies or the pie chart.

Then, use the function Replace to replace the inaccurate or wrong categories.