1   Subject

Dealing with outliers – Univariate tests with Tanagra (1.4.24 and later version).

The detection and the treatment of outliers (individuals with unusual values) is an important task of data preparation. Unusual values can mislead results of subsequent data analysis.

Outliers can be detected on one variable (a man with 158 years old) or on a combination of variables (a boy with 12 years old crosses the 100 yards in 10 seconds).

In this tutorial, we show how to use the **UNIVARIATE OUTLIER DETECTION** component. It is intended to univariate detection of outliers i.e. taking into account individually the variables.

The approaches implemented in the component come from the following website http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm. We use also an additional rule based on the x-sigma deviation from the mean of the variable.

The correspondence between x-sigma rule and the Tukey's box plot rule when we have a Gaussian distribution are displayed in the following chart (Figure 1).
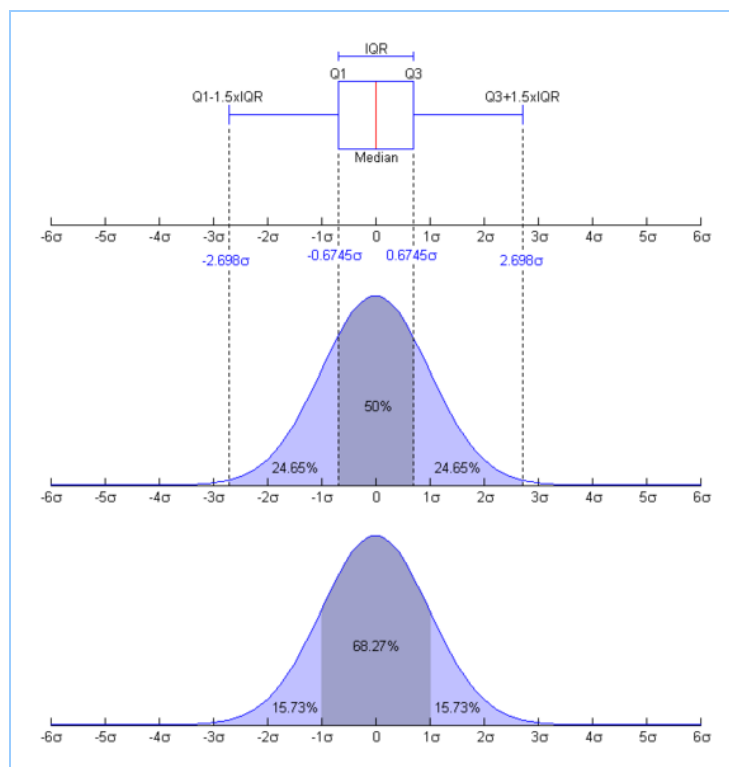

Figure 1 – Correspondence between the two rules of outliers detection for Gaussian distribution (http://en.wikipedia.org/wiki/Image:Boxplot_vs_PDF.png)

Even if these rules are efficient, we note in real problems that graphical approaches and/or descriptive statistics are often useful in many contexts. In fact, numerical methods are really interesting when we want to automatically deal with a large number of variables.
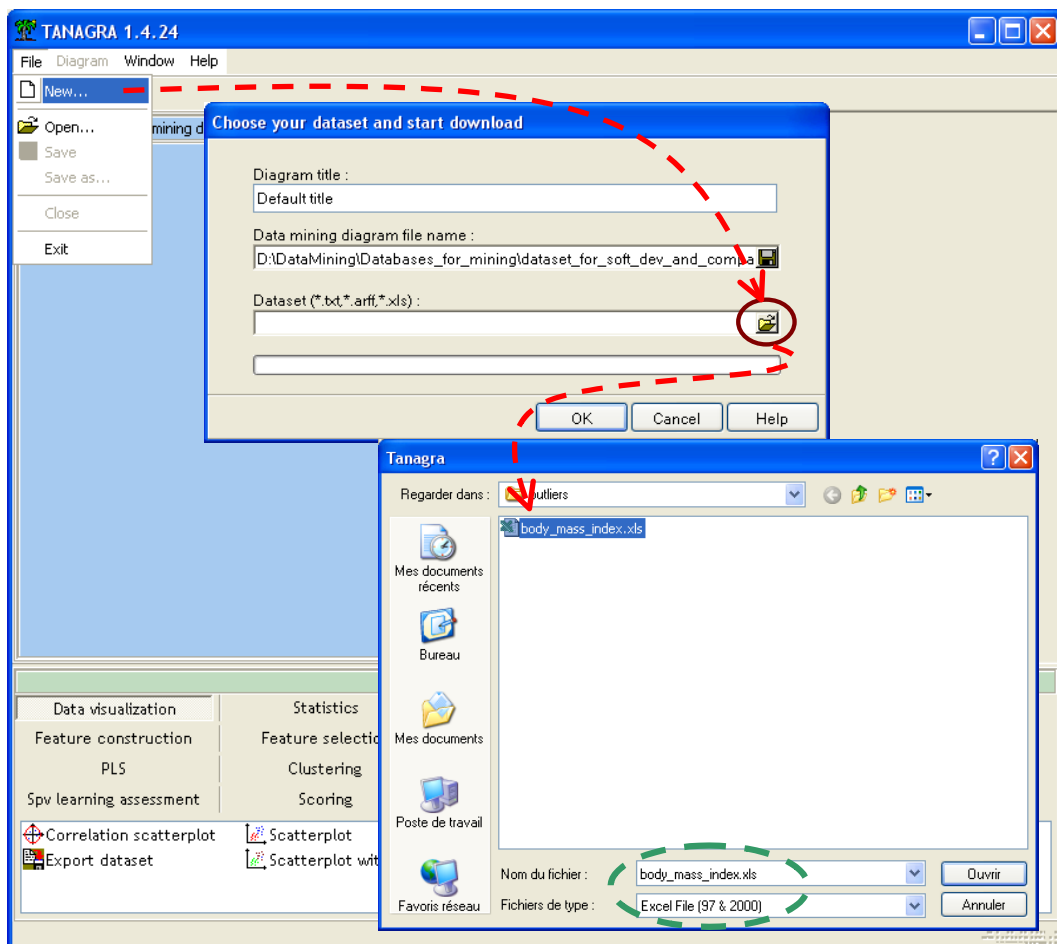
2   Dataset

The body_mass_index.xls[1] data file contains 50 examples and 3 variables. The variables are WEIGHT in KG, HEIGHT in meters and the BODY MASS index. We want to detect unusual values for each variable.

3   Outliers detection with TANAGRA

3.1   Creating a new diagram

We can directly import a XLS data file by creating a new diagram. This solution does not require the presence of the EXCEL software on the computer[2]. Caution: the data file has not to be currently edited; the dataset must be in the first sheet.

We launch Tanagra; we activate the FILE / NEW menu. Then we specify the name of the XLS file and the diagram filename.


The dataset is now available for the statistical analysis.

[1] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/body_mass_index.xls

[2] We can also open the file in the spreadsheet and send the dataset to Tanagra using an add-in. See http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/en_Tanagra_Excel_AddIn.pdf

---

3.2 Descriptive statistics

Descriptive statistics enables to describe shortly the main characteristics of a collection of observed data. First, we must define the INPUT variables. We use the DEFINE STATUS component.
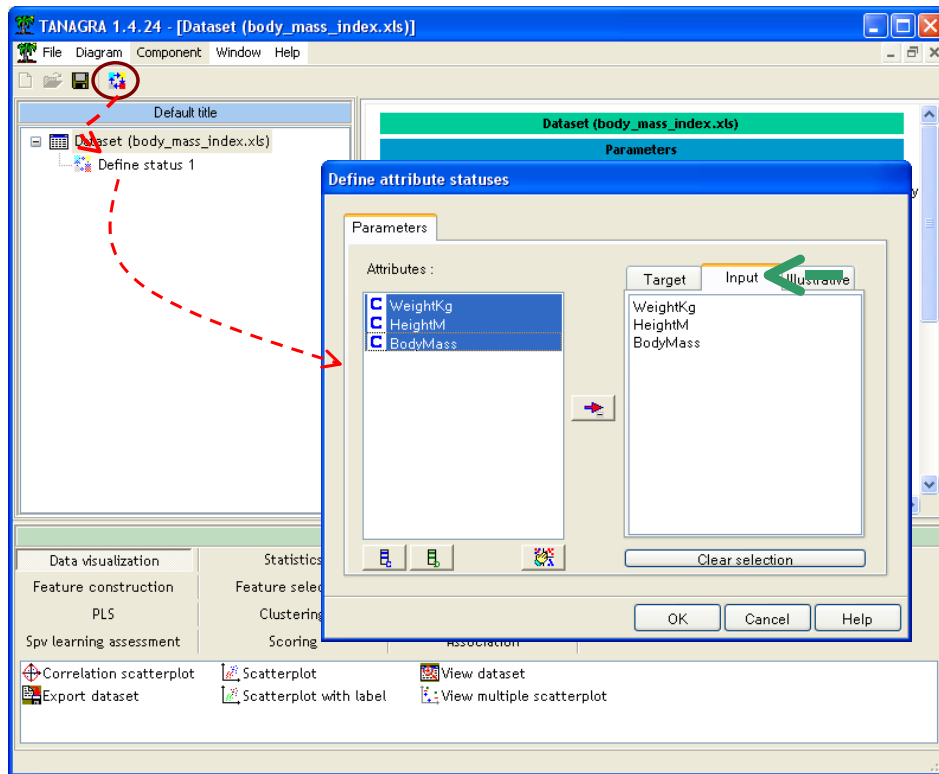
Then, we insert the MORE UNIVARIATE CONT STAT (STATISTICS tab) component into the diagram. Some numerical indicators and charts (frequency distribution chart) are computed.
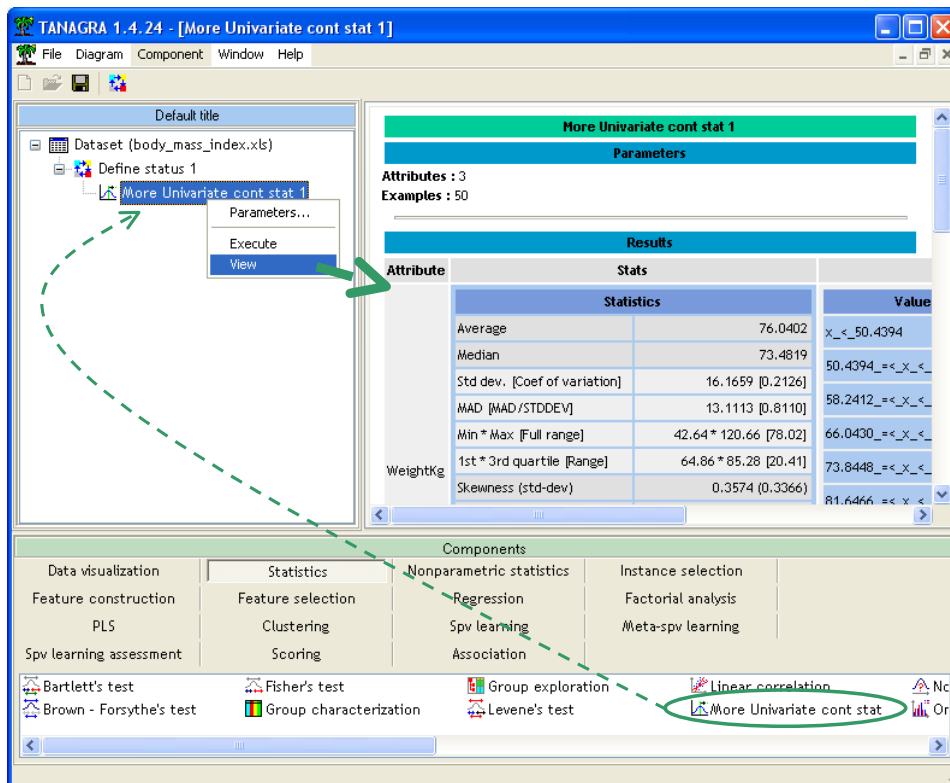
We summarized in the following table the indicators and their meanings.

Indicator	Description
Average	Mean
Median	Mean
Std.Dev. [Coef of variation]	Standard deviation and coefficient of variation (std.dev / average)
MAD [MAD / STDDEV]	Mean absolute deviation[3]. MAD / Std.Dev ratio
Min, Max [Full Range]	Minimum, maximum, range
1st * 3rd quartile [Range]	1er et $3^{ème}$ quartile ; inter quartile range
Skewness (std dev)	Skewness and its standard deviation
Kurtosis (std dev)	Kurtosis and its standard deviation

WEIGHTKG.

Attribute	Stats		Histogram			
	Statistics		**Values**	**Count**	**Percent**	**Histogram**
	Average	76.0402	x_<_50.4394	2	4.00%	
	Median	73.4819	50.4394_=<_x_<_58.2412	4	8.00%	
	Std dev. [Coef of variation]	16.1659 [0.2126]	58.2412_=<_x_<_66.0430	8	16.00%	
	MAD [MAD/STDDEV]	13.1113 [0.8110]	66.0430_=<_x_<_73.8448	11	22.00%	
WeightKg	Min * Max [Full range]	42.64 * 120.66 [78.02]	73.8448_=<_x_<_81.6466	7	14.00%	
	1st * 3rd quartile [Range]	64.86 * 85.28 [20.41]	81.6466_=<_x_<_89.4483	7	14.00%	
	Skewness (std-dev)	0.3574 (0.3366)	89.4483_=<_x_<_97.2501	8	16.00%	
	Kurtosis (std-dev)	0.1363 (0.6619)	97.2501_=<_x_<_105.0519	1	2.00%	
			105.0519_=<_x_<_112.8537	1	2.00%	
			x>=_112.8537	1	2.00%	

HEIGHTM.

	Statistics		Values	Count	Percent	Histogram
	Average	1.6581	x_<_1.4902	2	4.00%	
	Median	1.6510	1.4902_=<_x_<_1.5352	3	6.00%	
	Std dev. [Coef of variation]	0.1047 [0.0632]	1.5352_=<_x_<_1.5801	9	18.00%	
	MAD [MAD/STDDEV]	0.0901 [0.8608]	1.5801_=<_x_<_1.6251	8	16.00%	
	Min * Max [Full range]	1.45 * 1.89 [0.45]	1.6251_=<_x_<_1.6701	5	10.00%	
HeightM	1st * 3rd quartile [Range]	1.58 * 1.74 [0.17]	1.6701_=<_x_<_1.7150	5	10.00%	
	Skewness (std-dev)	0.0646 (0.3366)	1.7150_=<_x_<_1.7600	8	16.00%	
	Kurtosis (std-dev)	-0.8721 (0.6619)	1.7600_=<_x_<_1.8049	8	16.00%	
			1.8049_=<_x_<_1.8499	1	2.00%	
			x>=_1.8499	1	2.00%	

BODYMASS.

---

[3] http://en.wikipedia.org/wiki/Absolute_deviation

	Statistics		Values	Count	Percent	Histogram
BodyMass	Average	27.6806	x_<_20.7198	2	4.00%	
	Median	26.9761	20.7198_=<_x_<_23.8059	9	18.00%	
	Std dev. [Coef of variation]	5.8125 [0.2100]	23.8059_=<_x_<_26.8920	12	24.00%	
	MAD [MAD/STDDEV]	4.0471 [0.6963]	26.8920_=<_x_<_29.9782	15	30.00%	
	Min * Max [Full range]	17.63 * 48.49 [30.86]	29.9782_=<_x_<_33.0643	6	12.00%	
	1st * 3rd quartile [Range]	24.02 * 29.65 [5.63]	33.0643_=<_x_<_36.1504	4	8.00%	
	Skewness (std-dev)	1.5480 (0.3366)	36.1504_=<_x_<_39.2366	0	0.00%	
	Kurtosis (std-dev)	4.3365 (0.6619)	39.2366_=<_x_<_42.3227	0	0.00%	
			42.3227_=<_x_<_45.4088	0	0.00%	
			x>=_45.4088	2	4.00%	

Two examples seem different to others for each variable. But we do not know if these are the same ones for all the variables.

3.3 Scatter plot

We can visualize the examples in a scatter plot. It enables us to detect the deviation of the examples in a bivariate way i.e. taking into account the interaction between variables.

We insert the SCATTERPLOT WITH LABEL (DATA VISUALIZATION tab) component into the diagram. We set WEIGHTKG for the horizontal axis, HEIGHTM for the vertical axis.
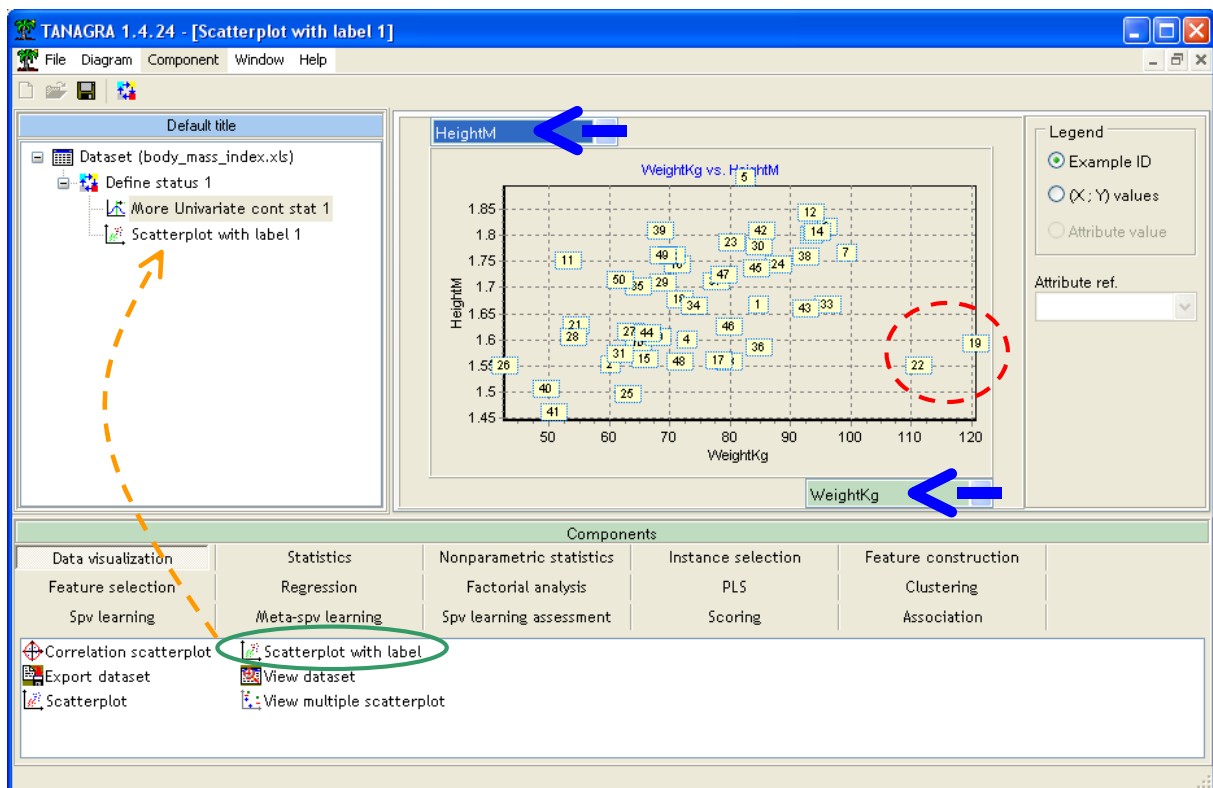
The examples 19 and 22 seem abnormal compared with the main pattern of the points. The two examples with the highest values of HEIGHTM are not the examples with the highest values of WEITHKG.

We set now WEIGHTKG and BODYMASS in the scatter plot.
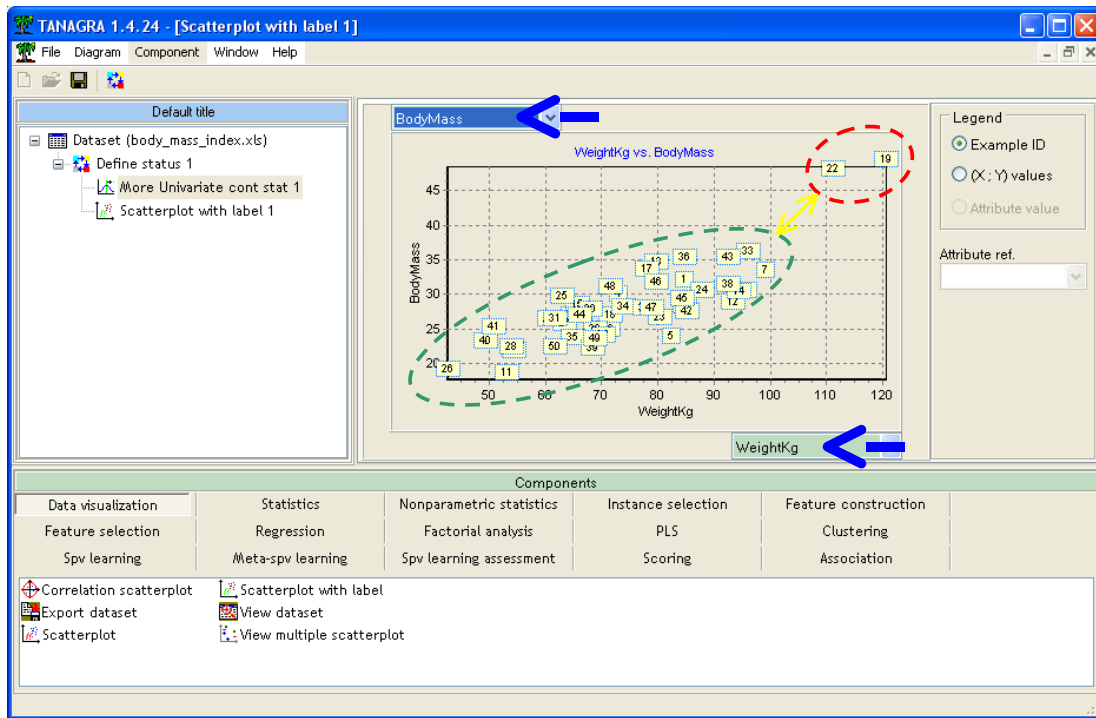
The two individuals (n°19 and n°22) seem really overweighted.

3.4 Automatic detection

The UNIVARIATE OUTLIER DETECTION component tries to detect the examples that are in an abnormal distance from other values. It uses the various criteria enumerated here http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm. We insert the component into the diagram. We activate the contextual PARAMETERS menu.
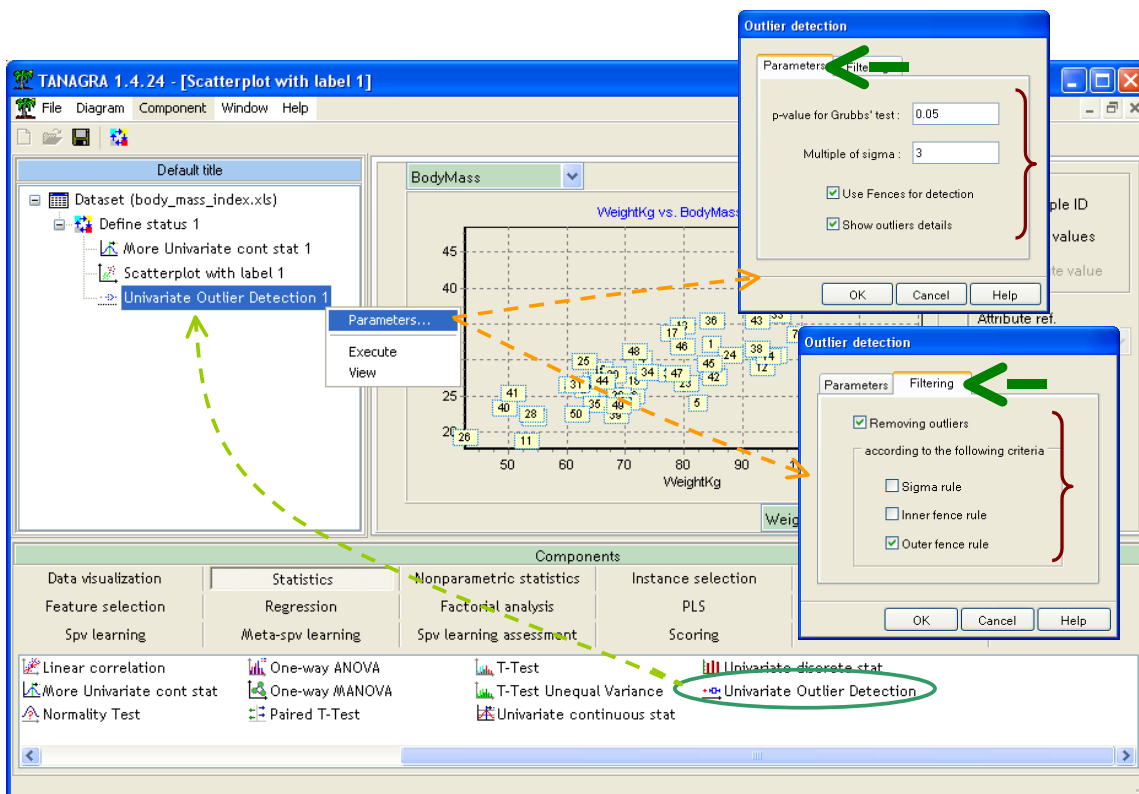
In PARAMETERS tab, we choose to display the details of detected outliers in the report. In FILTERING tab, we choose to remove the outliers according the last criterion i.e. the OUTER FENCE rule (see section **Erreur ! Source du renvoi introuvable.**). We validate these parameters and we activate the VIEW contextual menu to access the results.
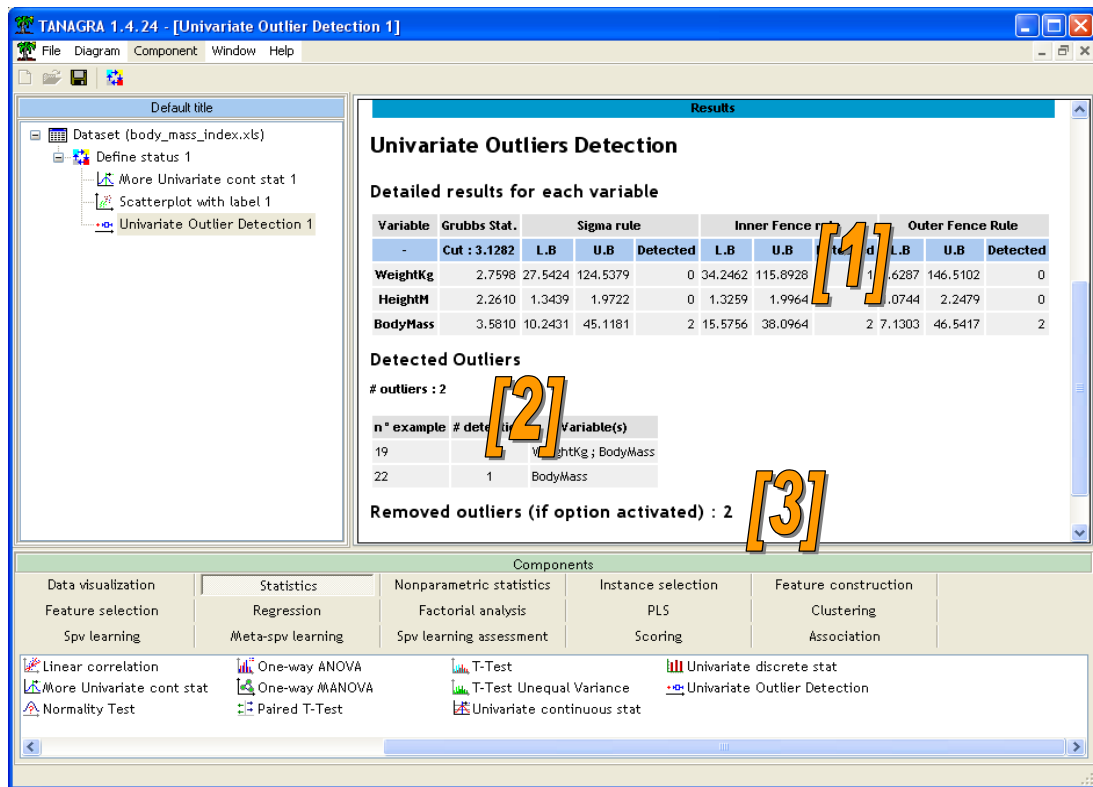
In the first part of the report [1], we observe the limit values and the number of atypical observations for each criterion.

Detailed results for each variable

Variable	Grubbs Stat.	Sigma rule			Inner Fence rule			Outer Fence Rule		
-	Cut : 3.1282	L.B	U.B	Detected	L.B	U.B	Detected	L.B	U.B	Detected
WeightKg	2.7598	27.5424	124.5379	0	34.2462	115.8928	1	3.6287	146.5102	0
HeightM	2.2610	1.3439	1.9722	0	1.3259	1.9964	0	1.0744	2.2479	0
BodyMass	3.5810	10.2431	45.1181	2	15.5756	38.0964	2	7.1303	46.5417	2

- For the Grubbs test, at the significance level of 5%, the most extreme value of BODYMASS is an outlier.

- According to the 3-sigmas criterion, there are two abnormal values for BODYMASS.

- According to the INNER FENCE criterion, there is 1 outlier for WEIGHTKG, 2 for BODYMASS.

- According to the OUTER FENCE criterion, we have the same results as 3-sigmas.

In the second part of the report [2], the details about abnormal examples, according to the variables, are displayed.

The example n°19 is abnormal according two variables; n°22 according BODYMASS only.

Finally, in the third part of the report [3], we observe that 2 individuals are removed from the dataset, according the OUTER FENCE criterion specified above.

3.5 Descriptive statistics (again)

To evaluate the influence of the removed examples (n°19 and n°22), we compute again the descriptive statistics, especially the mean, without these examples. Then we compare the results with the previous values. We insert the UNIVARIATE CONTINUOUS STAT (STATISTICS tab) into the diagram.
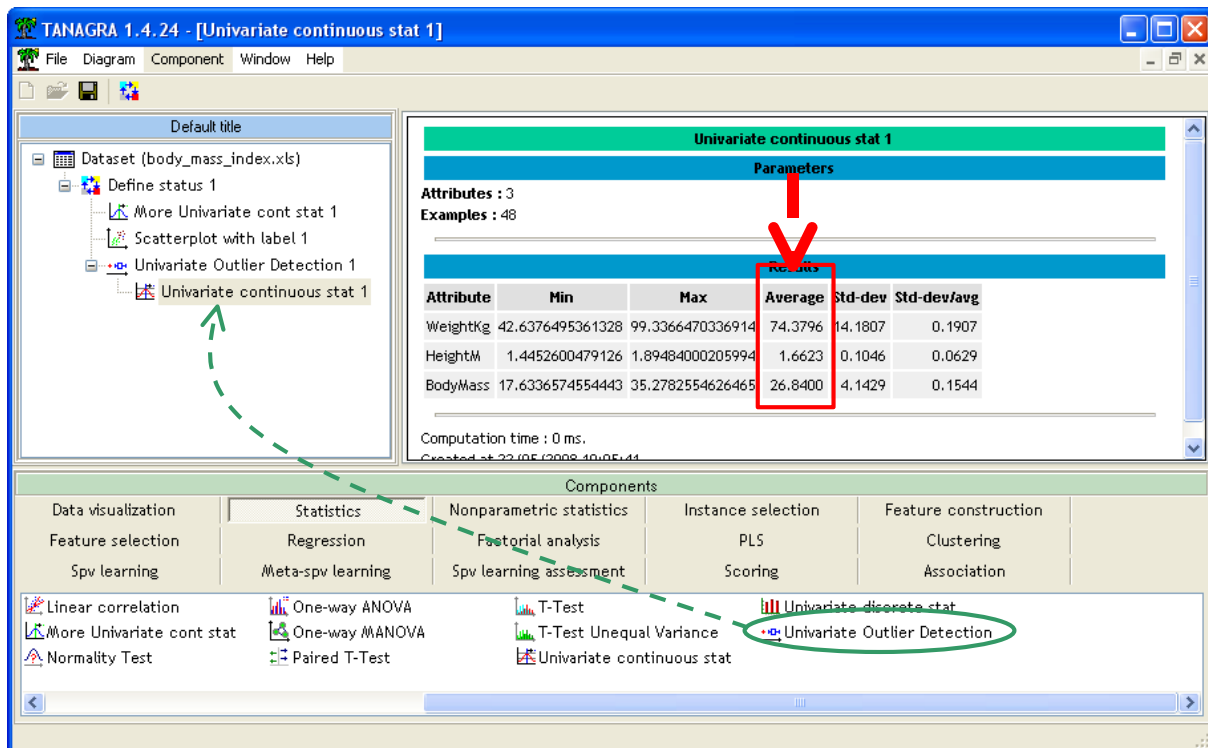
The deviation is particularly high for BODYMASS.

Variable	Mean for 50 examples	Mean for 48 ex. (without n°19 and n°22)	Deviation (en %)
WEIGHTKG	76.0402	74.3796	+2.23 %
HEIGHTM	1.6581	1.6623	-0.25 %
MODYMASS	27.6806	26.8400	+3.13 %

4   Conclusion – Outliers treatment

We choose to delete the outliers in our tutorial. This is not generally the best strategy. Abnormal examples can contribute highly to the data exploration. There are various approaches which enable to integrate them in the analysis without their disadvantages (see for instance http://cc.uoregon.edu/cnews/spring2000/outliers.html).

Another key point of this tutorial (and the studied component) is the utilization of univariate strategy. It does not take into account the interaction among the variables. Sometimes, an example is not abnormal on all the variables (individually treated), but abnormal on combination of variables. We must use more sophisticated method in this context.