

## Subject

### Descriptive statistics with TANAGRA.

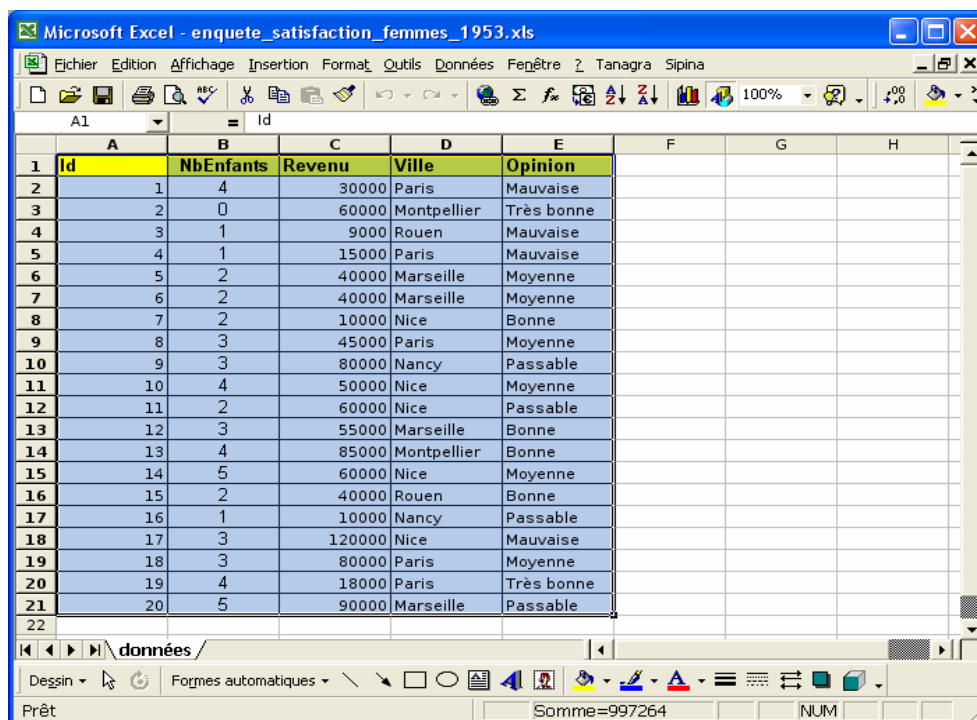
The aim of descriptive statistics is to describe the main features of a collection of data in quantitative terms<sup>1</sup>. The visualization of the whole data table is seldom useful. It is preferable to summarize the characteristics of the data with some selected numerical indicators.

In this tutorial, we distinguish two kinds of descriptive approaches: the univariate tools which summarize the characteristics of a variable individually; the bivariate tools which characterize the association between two variables. According to the type of the variables (categorical or continuous), we use different indicators.

## Creating a diagram

### Dataset

The dataset comes from a consumer survey about a product for baby. There are 20 observations, ID is an identifier. The variables are interval (REVENU – INCOME), categorical (VILLE – TOWN) or ordinal (OPINION – JUDGEMENT). The number of children is more complicated. It is both a continuous variable (the difference between two values has a signification) and an ordinal variable (the number of values is limited).

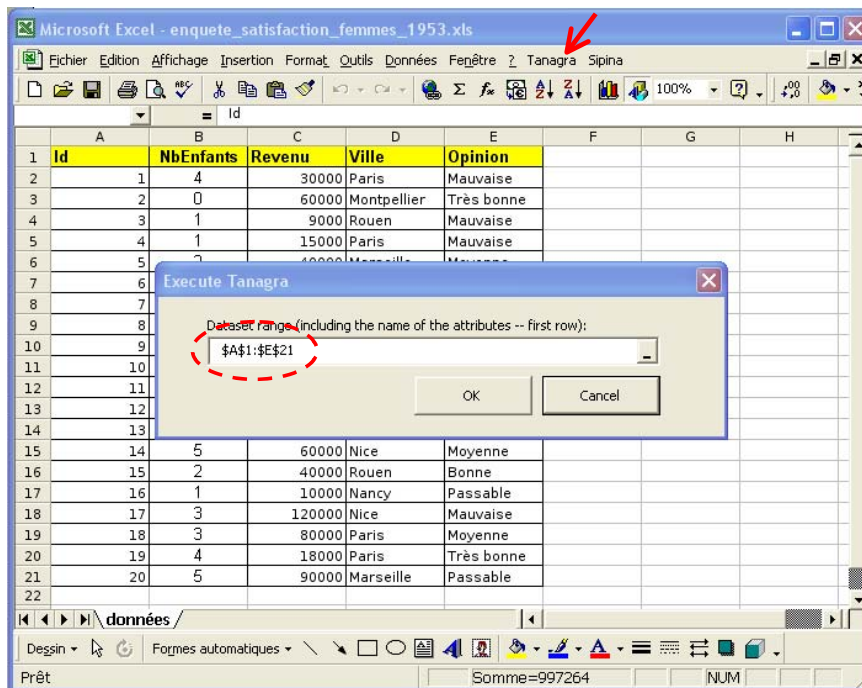


	A	B	C	D	E	F	G	H
1	Id	NbEnfants	Revenu	Ville	Opinion			
2	1	4	30000	Paris	Mauvaise			
3	2	0	60000	Montpellier	Très bonne			
4	3	1	9000	Rouen	Mauvaise			
5	4	1	15000	Paris	Mauvaise			
6	5	2	40000	Marseille	Moyenne			
7	6	2	40000	Marseille	Moyenne			
8	7	2	10000	Nice	Bonne			
9	8	3	45000	Paris	Moyenne			
10	9	3	80000	Nancy	Passable			
11	10	4	50000	Nice	Moyenne			
12	11	2	60000	Nice	Passable			
13	12	3	55000	Marseille	Bonne			
14	13	4	85000	Montpellier	Bonne			
15	14	5	60000	Nice	Moyenne			
16	15	2	40000	Rouen	Bonne			
17	16	1	10000	Nancy	Passable			
18	17	3	120000	Nice	Mauvaise			
19	18	3	80000	Paris	Moyenne			
20	19	4	18000	Paris	Très bonne			
21	20	5	90000	Marseille	Passable			
22								

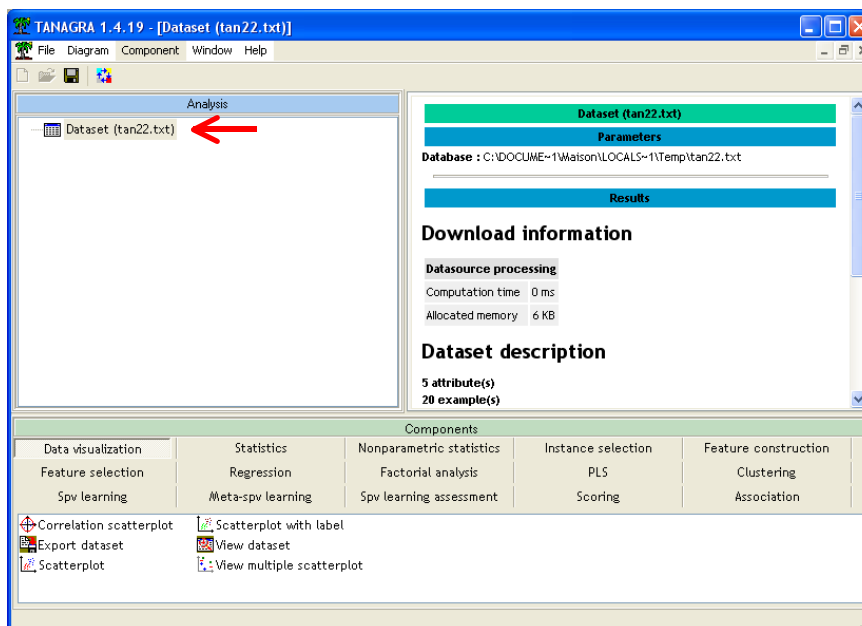
### Importing the data file

First, we must create a new diagram and import the data file « enquete\_satisfaction\_femmes\_1953.xls ». The easiest way is to open the file in Excel. We select the range of cells, including the first row corresponding to the name of the variables, and we click on the TANAGRA / EXECUTE TANAGRA menu<sup>2</sup>.

<sup>1</sup> [http://en.wikipedia.org/wiki/Descriptive\\_statistics](http://en.wikipedia.org/wiki/Descriptive_statistics)



TANAGRA is automatically launched. A new diagram is created and the dataset is available for the analysis.



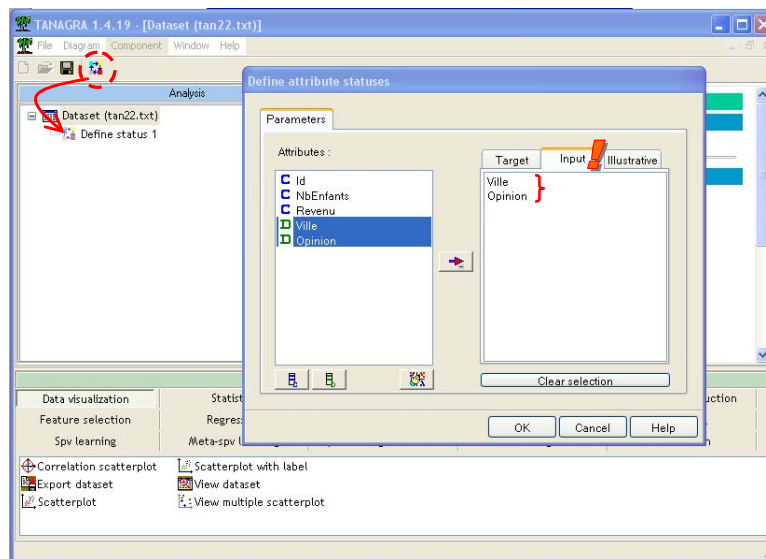
## Analyzing Categorical Variables

### Univariate statistics

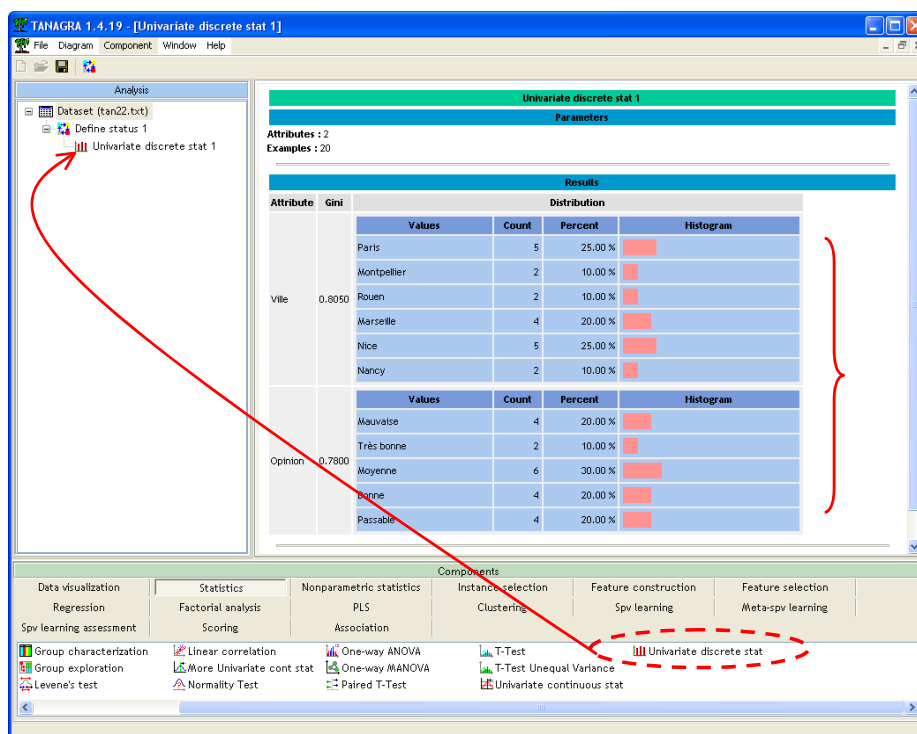
The first tool for characterizing the distribution of a categorical attribute is the absolute frequency. For each value of the variable, we count the number of examples. We can also compute the percentage of examples.

<sup>2</sup> This menu is available after we install the TANAGRA.XLA add-on into Excel. The procedure is described here: <http://data-mining-tutorials.blogspot.com/2008/10/excel-file-handling-using-add-in.html>

We insert the DEFINE STATUS component into the diagram using the shortcut into the tool bar. We set VILLE and OPINION as INPUT attributes.



We add now the UNIVARIATE DISCRETE STAT (STATISTICS tab) into the diagram. We click on the VIEW menu in order to obtain the results.



We note for instance that 25% of the respondents come from Paris; 10% from Montpellier; etc.

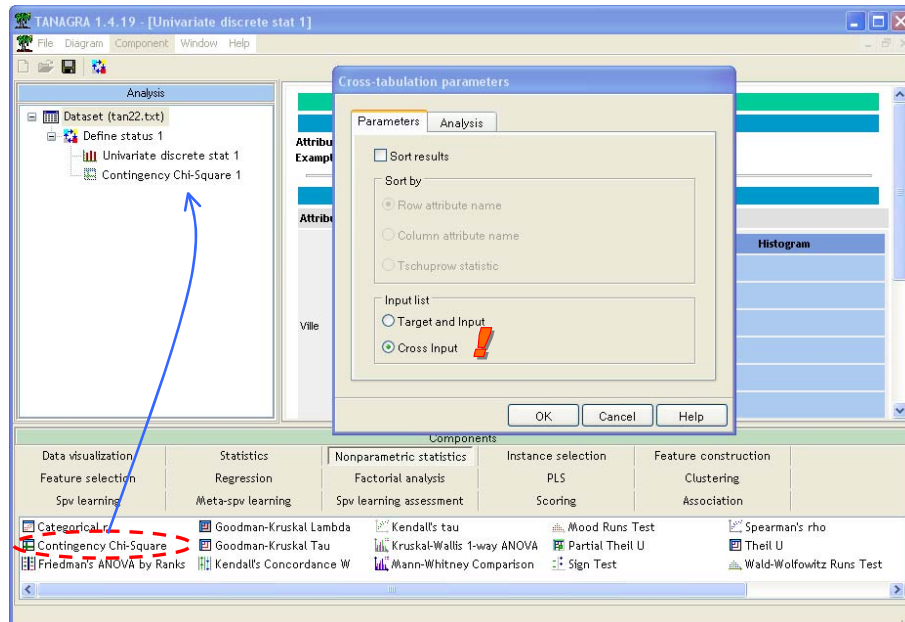
The Gini index helps to locate the concentration of observations on some values. If it is close to 0, the observations are concentrated on a few values of the variable.

### Bivariate descriptive statistics for categorical variables

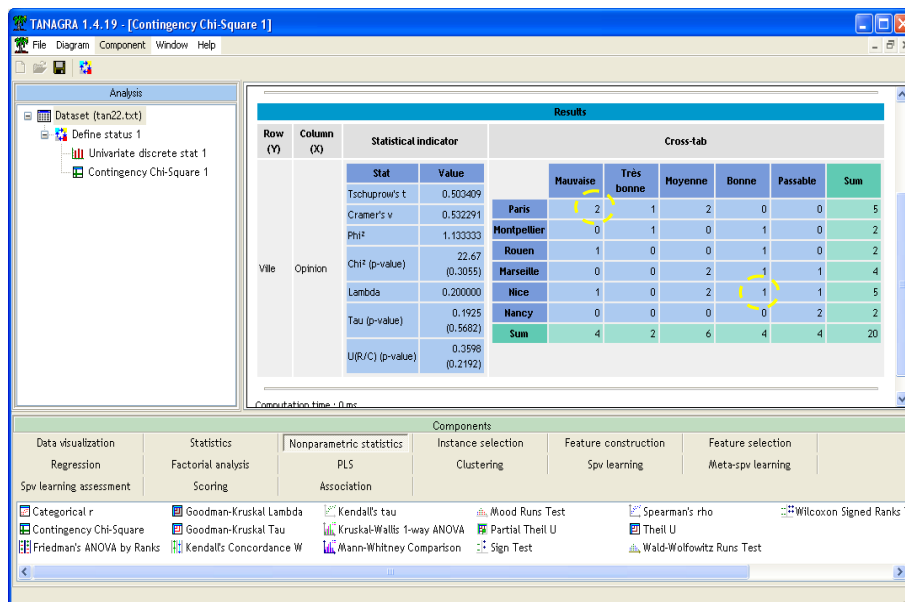
We want to characterize now the distribution of OPINION according to VILLE. We create a contingency table.

## Absolute frequencies

After the DEFINE STATUS component into the diagram, we add the component CONTINGENCY CHI-SQUARE (NONPARAMETRIC STATISTICS tab). We click on the PARAMETERS contextual menu. The variables are listed in the INPUT list i.e. the component computes the contingency table for each pair of variables.



We click on the VIEW menu, we obtain the following results.



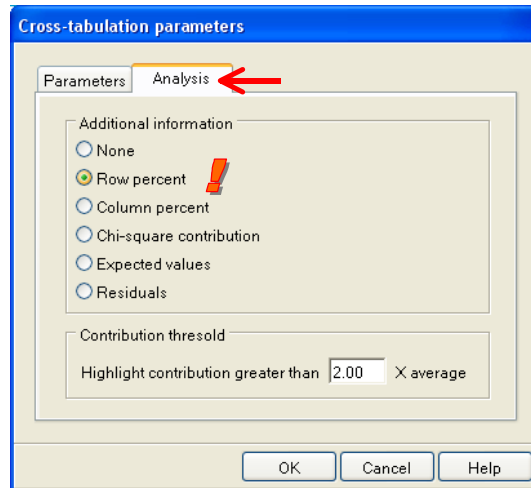
We observe for instance that two individuals from Paris have a bad (mauvaise) opinion about the product; 1 individual from Nice has a good (bonne) opinion; etc.

Various numerical indicators try to quantify the strength of the association between the variables (see <http://www2.chass.ncsu.edu/garson/PA765/assocnominal.htm> or <http://en.wikipedia.org/wiki/Crosstab>).

## Percentage by row or percentage by column

Often, the absolute frequencies are not informative. It is better to compute relative frequencies, by row or by column. About the opinion according the town, it is more suitable to dividing the absolute values by the margin. The values are then comparable.

We click on the PARAMETERS menu. We select the ROW PERCENT option.



By clicking on the VIEW menu, we obtain:

Row (Y)	Column (X)	Statistical indicator		Cross-tab						
		Stat	Value	Mauvaise	Très bonne	Moyenne	Bonne	Passable	Sum	
Ville	Opinion	Tschuprow's t	0.503409							
		Cramer's v	0.532291							
		Phi²	1.133333							
		Chi² (p-value)	22.67 (0.3055)							
		Lambda	0.200000							
		Tau (p-value)	0.1925 (0.5682)							
		U(R/C) (p-value)	0.3598 (0.2192)							
		Paris		2	1	2	0	0	5	
		Montpellier		0	1	0	1	0	2	
		Rouen		1	0	0	1	0	2	
Marseille		0	0	2	1	1	4			
Nice		1	0	2	1	1	5			
Nancy		0	0	0	0	2	2			
Sum		4	2	6	4	4	20			
		20%	10%	30%	20%	20%	100%			

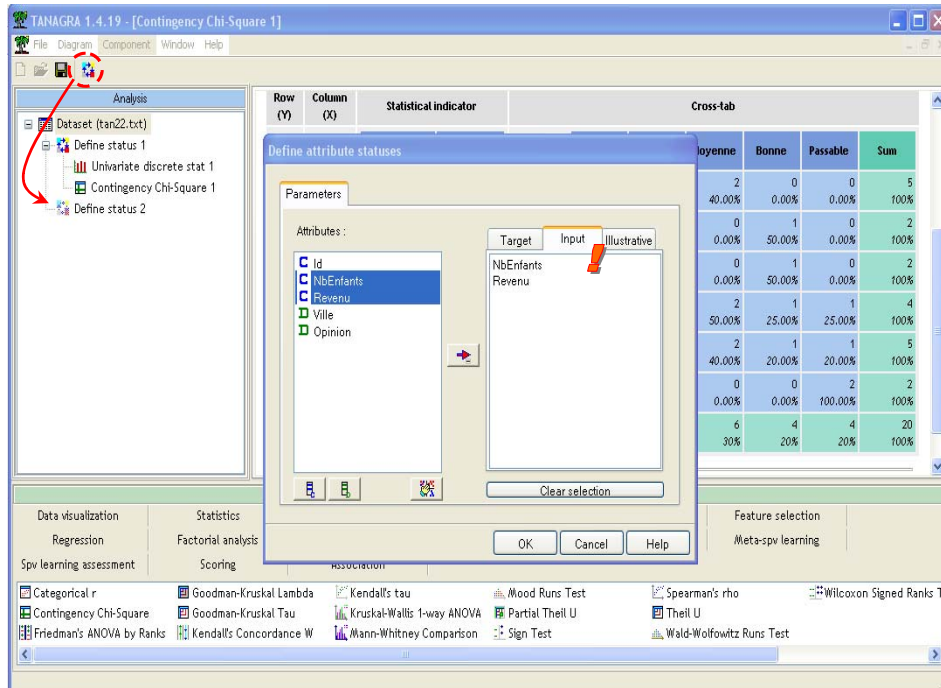
For instance, the absolute frequency of good opinion is the same in Montpellier and Nice. When we divide the value by the number of instances in each town, we note that the proportions are not the same.

# Continuous variables

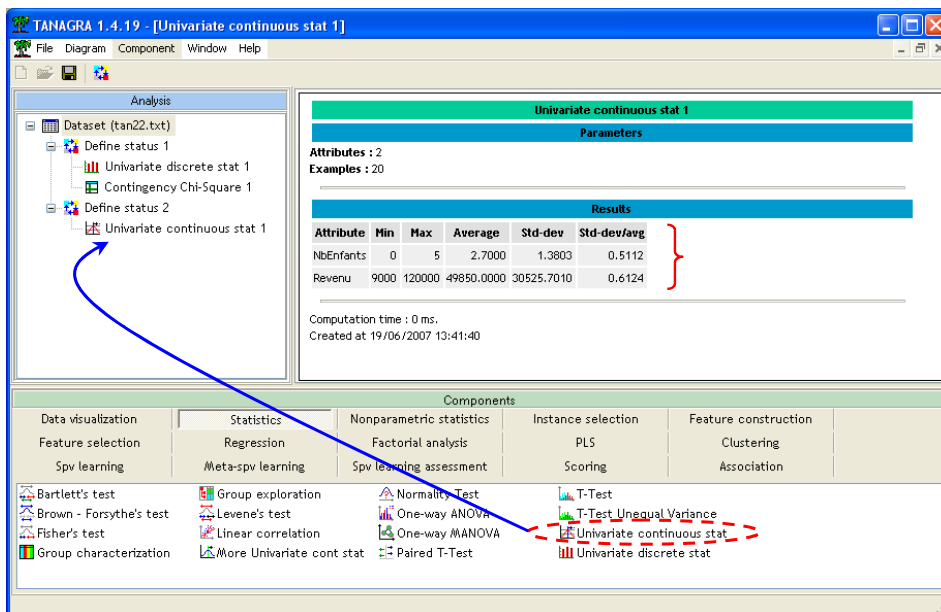
## Univariate statistics

Two components are dedicated to univariate descriptive statistics for continuous variable.

We insert again the DEFINE STATUS component. We set as INPUT both NBENFANTS (number of children) and REVENU. Computing descriptive statistics on the ID column is not really useful.

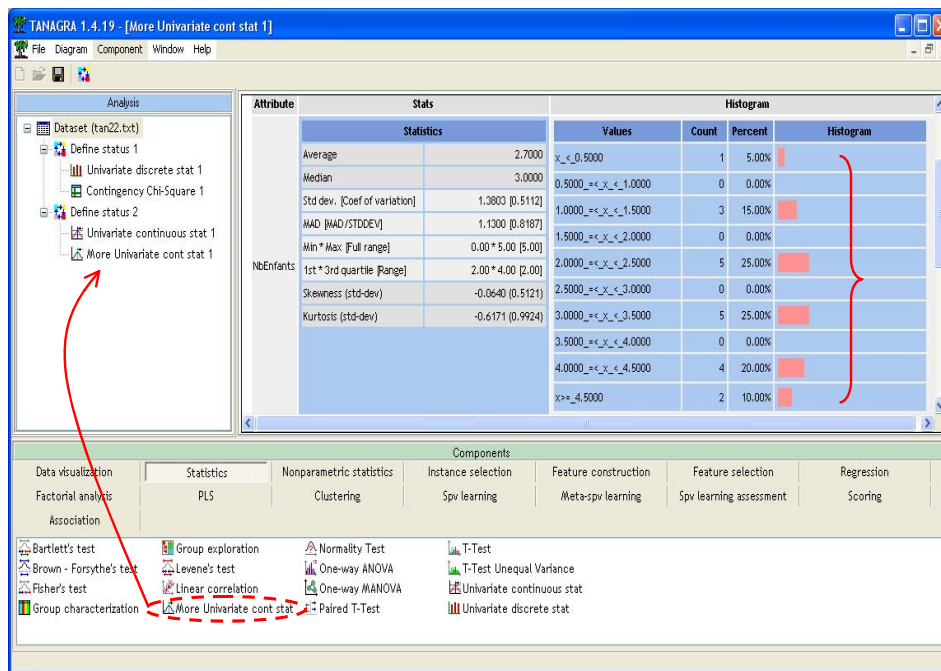


Then we add the UNIVARIATE CONTINUOUS STAT (STATISTICS tab) component. We click on VIEW menu.



This component computes simple indicators such as mean, standard deviation, etc. It is useful when we want to obtain quickly an overview on the characteristics of a large number of variables.

A second tool gives more detailed information. We add the MORE UNIVARIATE CONTINUOUS STAT component (STATISTICS tab). We obtain the following results.

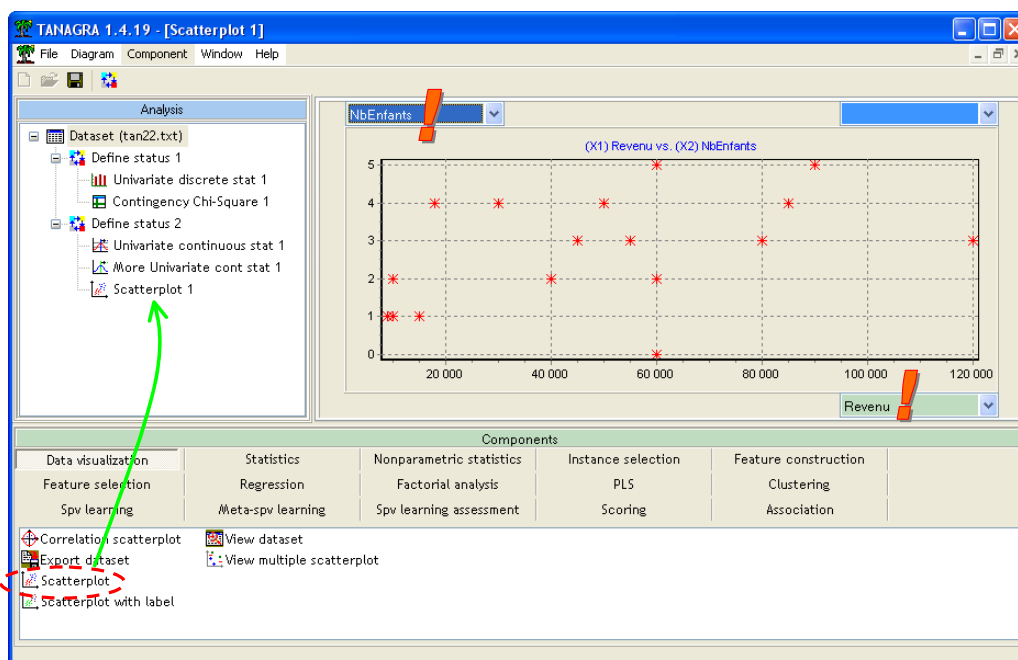


Various indicators and additional information about the shape of the distribution are supplied.

## Bivariate statistics for continuous variables

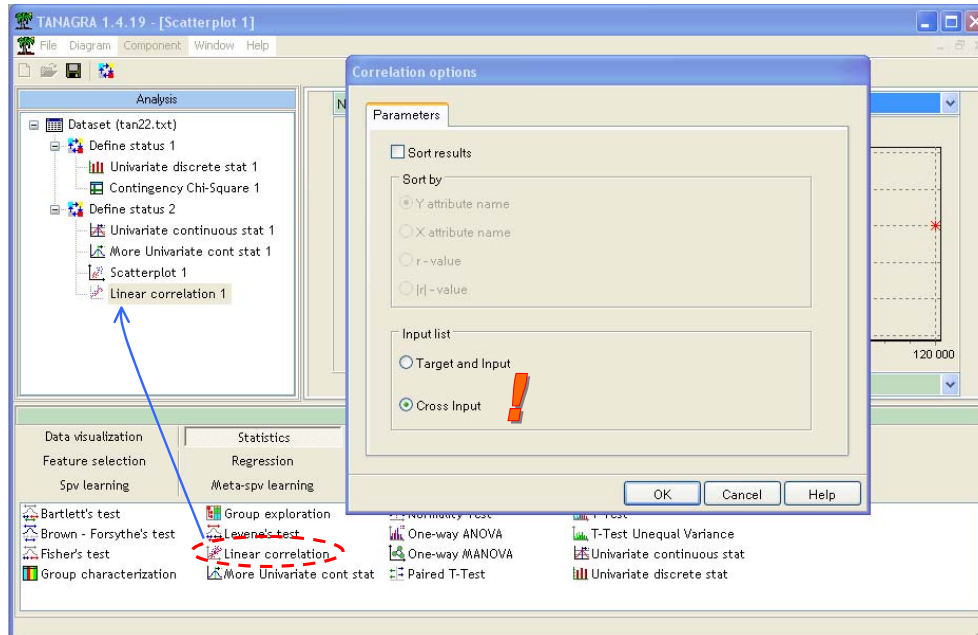
### Scatter plot

Unlike categorical variables, some graphical tools allow to evaluate the association between continuous variables. We insert the SCATTERPLOT component (VISUALIZATION tab). We obtain a scatter plot. We can modify interactively the variables in the horizontal and the vertical axis. The size of the points can be modified.

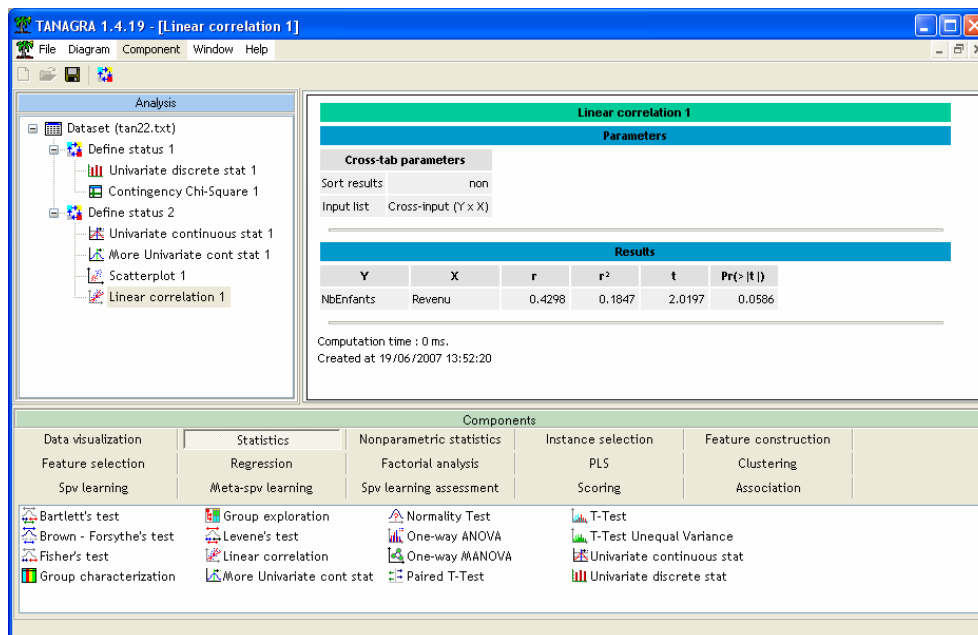


## Numerical indicators

We can characterize also the association with some numerical indicators. It is especially useful when we deal with a large number of variables. For instance, the LINEAR CORRELATION (STATISTICS tab) component allows to compute the correlation coefficient between the variables. In the settings dialog box, we indicate that the variables are listed in the INPUT section of the preceding DEFINE STATUS component.



We obtain the following result.



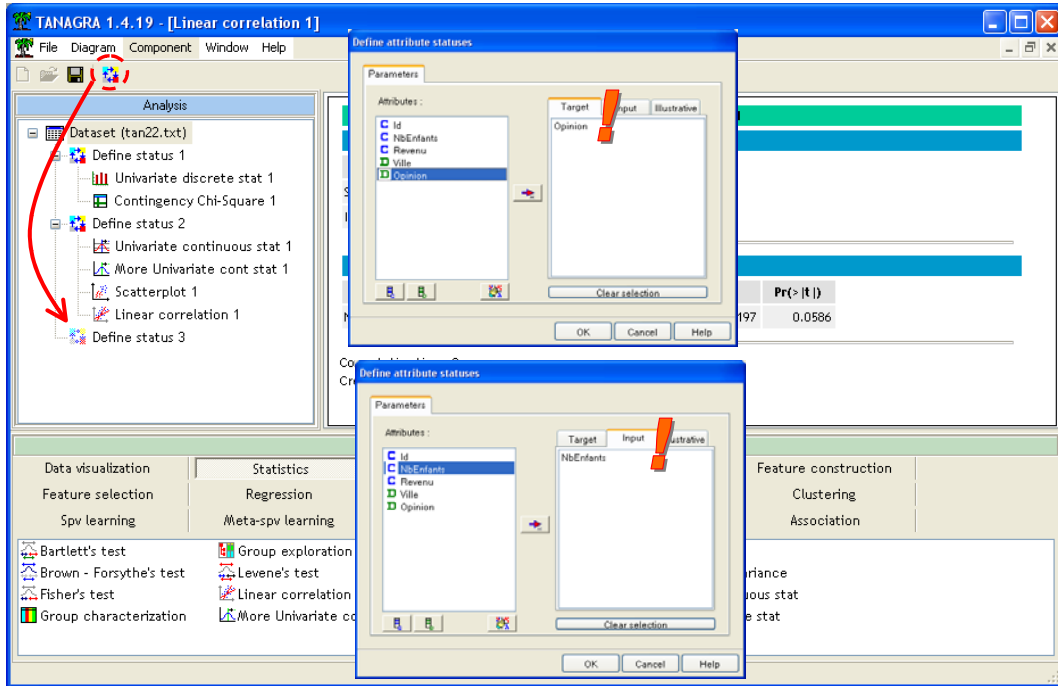
If the association is monotonic but not linear, we can use nonparametric statistics like SPEARMAN or KENDALL indicators (see NONPARAMETRIC STATISTICS tab).



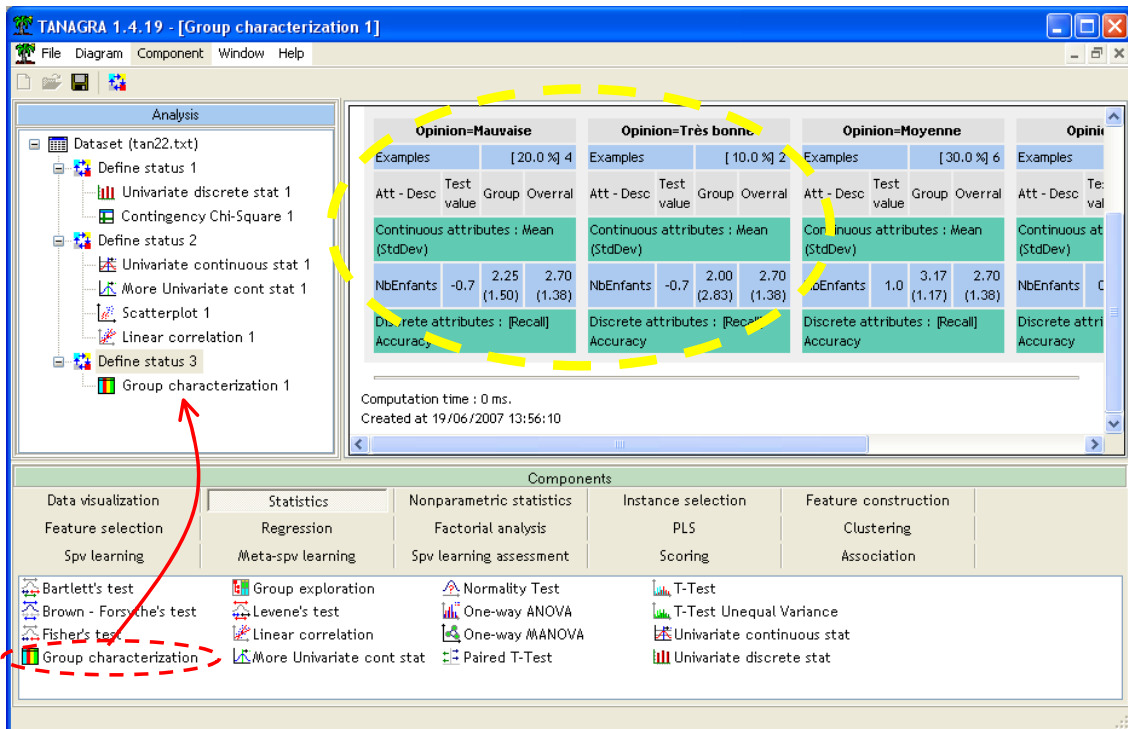
## Association between categorical and continuous variables

We can also compute the association between categorical and continuous variables. For instance, is the OPINION different according to the number of children?

We insert the DEFINE STATUS component into the diagram. We set OPINION as TARGET, NBENFANTS as INPUT.



Then we add the GROUP CHARACTERIZATION component (STATISTICS tab).



---

We find that women with a definite opinion, good or bad, have fewer children on average (conditional average is respectively 2.25 and 2.0, whereas the global average is 2.7).

## Conclusion

When we present the results of data mining study, basic indicators, tables and graphs, are often at least as relevant as the complicated statistical methods, fairly obscure for non-specialists. For this reason, descriptive statistics have always a large place in the reports.