

Subject

Measures of association for nominal variables.

To measure the association between two continuous variables, we generally use the correlation coefficient. Its drawbacks and its qualities are well known.

When we want to characterize the association for nominal variables, the correlation coefficient is not suitable. We must use other indicators. The most widespread is certainly the chi-square test, it enables to evaluate the absence of relation. We see in this tutorial that other measures are available. We show how to use them with TANAGRA.

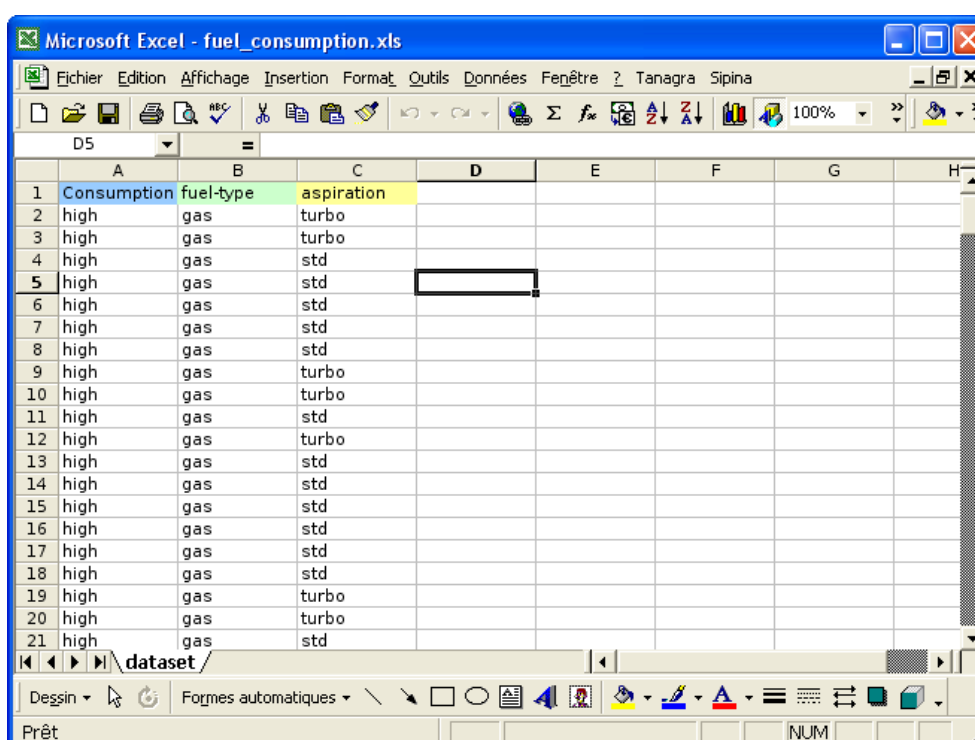
Descriptions of the measures depicted in this tutorial are available on the following website:

- http://www.georgetown.edu/faculty/ballc/webtools/web_chi_tut.html
- <http://v8doc.sas.com/sashtml/stat/chap28/sect20.htm>
- <http://www2.chass.ncsu.edu/garson/PA765/assocnominal.htm>

Dataset

In the FUEL_CONSUMPTION.XLS dataset, 205 cars are described by their CONSUMPTION, FUEL_TYPE and ASPIRATION.

We show here the 20 first examples.



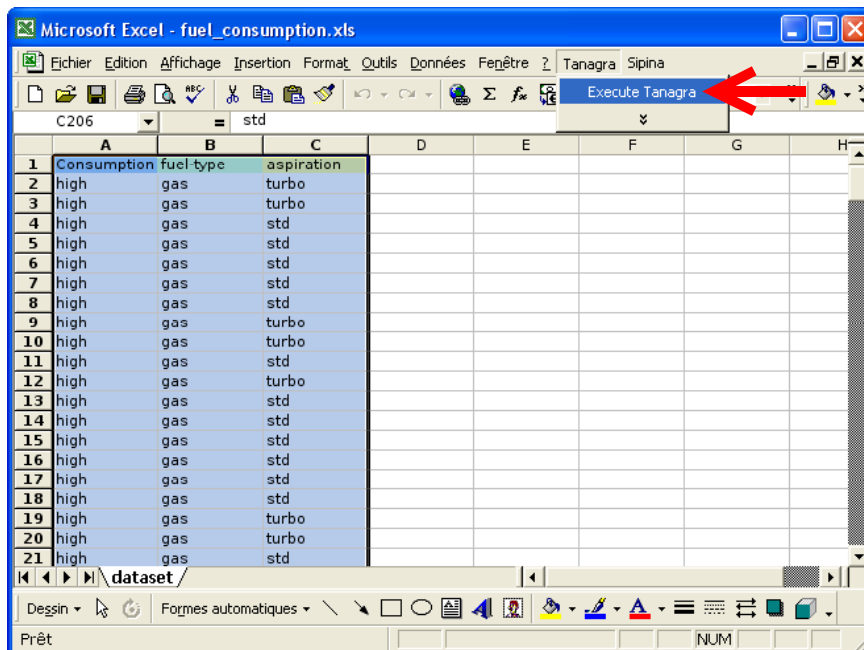
	A	B	C	D	E	F	G	H
1	Consumption	fuel-type	aspiration					
2	high	gas	turbo					
3	high	gas	turbo					
4	high	gas	std					
5	high	gas	std					
6	high	gas	std					
7	high	gas	std					
8	high	gas	std					
9	high	gas	turbo					
10	high	gas	turbo					
11	high	gas	std					
12	high	gas	turbo					
13	high	gas	std					
14	high	gas	std					
15	high	gas	std					
16	high	gas	std					
17	high	gas	std					
18	high	gas	std					
19	high	gas	turbo					
20	high	gas	turbo					
21	high	gas	std					

The goal is to measure the strength of the relation between ACCELERATION and the other variables.

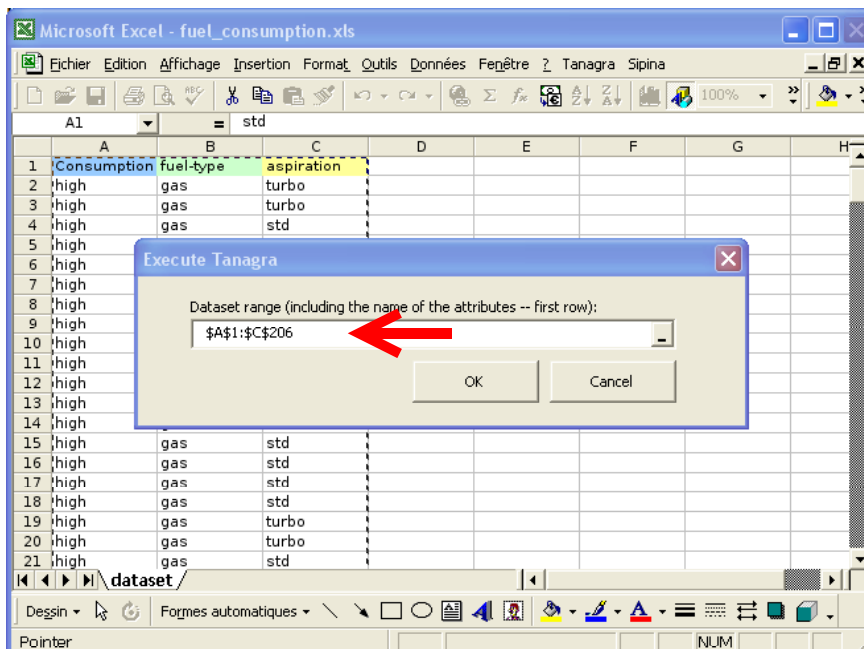
Chi Square test of independence

Create a new diagram

The simplest way in order to create a diagram is to load the dataset in the EXCEL spreadsheet. Then, we select the data range and we click on the menu TANAGRA/EXECUTE TANAGRA¹.

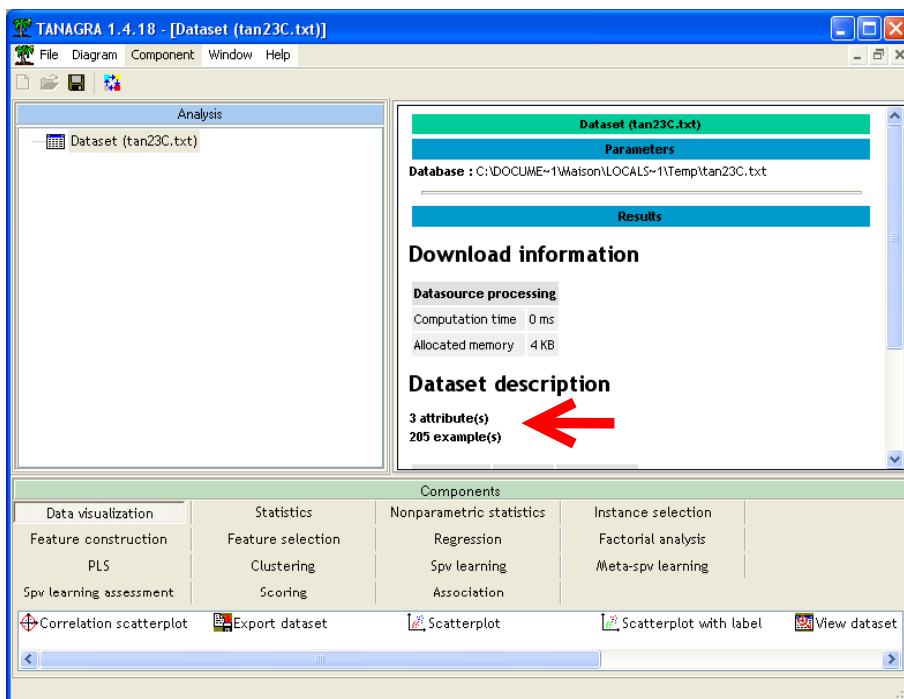


We check the range selection and we click on OK.



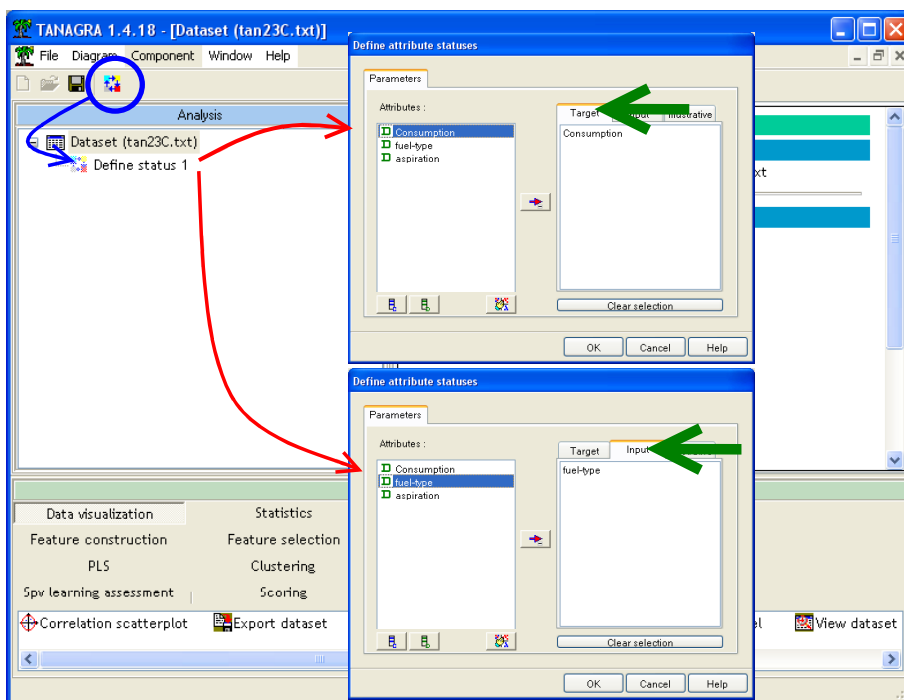
TANAGRA is automatically executed. We check that 3 variables and 205 examples are processed.

¹ The EXCEL add-in TANAGRA.XLA is available since the version 1.4.11. See the tutorial on the web site for the installation of this add-in in your spreadsheet.

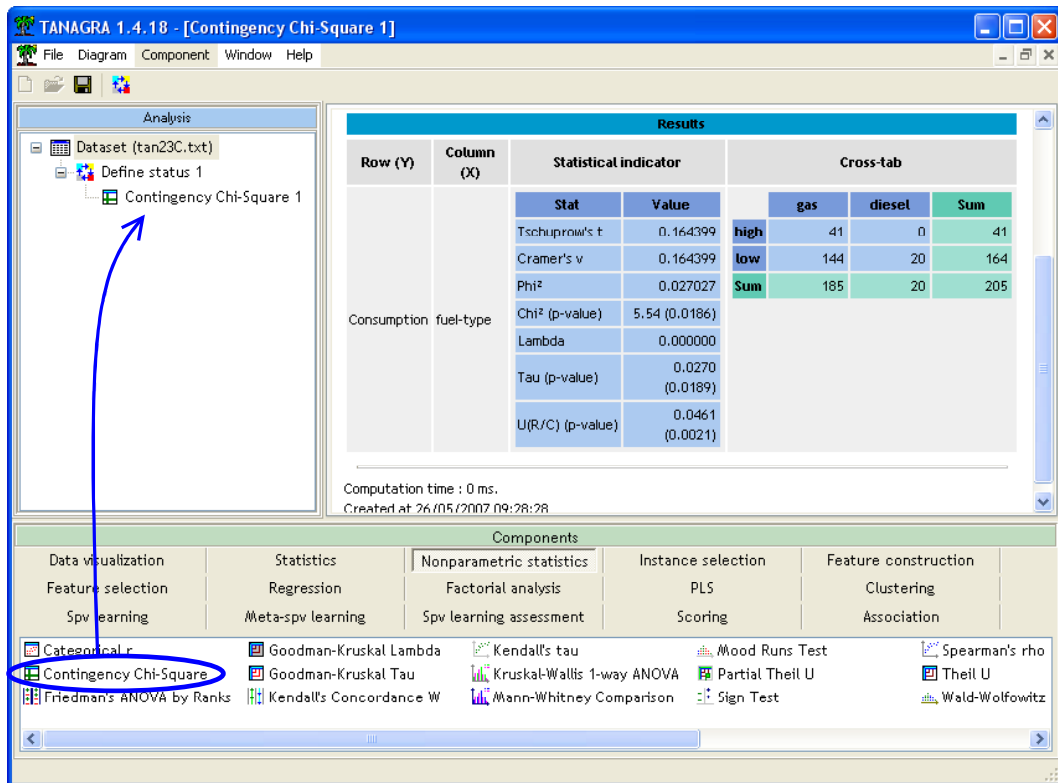


Crosstabulation table and chi square statistic

We add the DEFINE STATUS component by using the short cut into the toolbar. First, we want to analyze the relation between CONSUMPTION (TARGET) and FULE-TYPE (INPUT).



We add the component CONTINGENCY CHI-SQUARE (NONPARAMETRIC tab) into the diagram and we click on the VIEW menu.

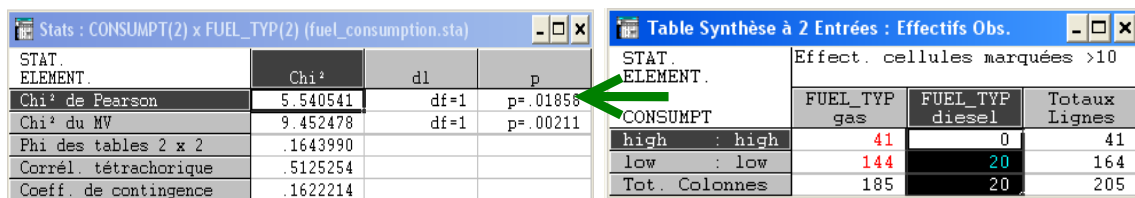


Crosstabulation table. CROSS-TAB is a contingency table. It allows examining the frequencies of observations that belong to each combination of categories of the variables. We note that there are 205 cars in the dataset, 20 of them use diesel fuel-type. Among these “FUEL-TYPE = diesel” cars, all have “CONSUMPTION = low”.

Chi-Square statistic. The Pearson CHI-SQUARE (CHI-2) is the most common test for significance of the relationship for nominal variables. This measure is based on the fact that we can compute the *expected* frequencies in the contingency table (i.e., frequencies that we would *expect* if there was no relationship between the variables). Chi-square measures the deviation of the observed frequencies to the expected frequencies. It is 5.54 in our dataset.

The *p-value* of the test allows us to determine if you must accept or reject the null hypothesis. We can compare this value with a user-predefined significance level (p-level) of the test (e.g. 5% level in the most cases). In our dataset, the p-value is 0.0189. We reject the null hypothesis (reject "absence of relation") for a 5%-level.

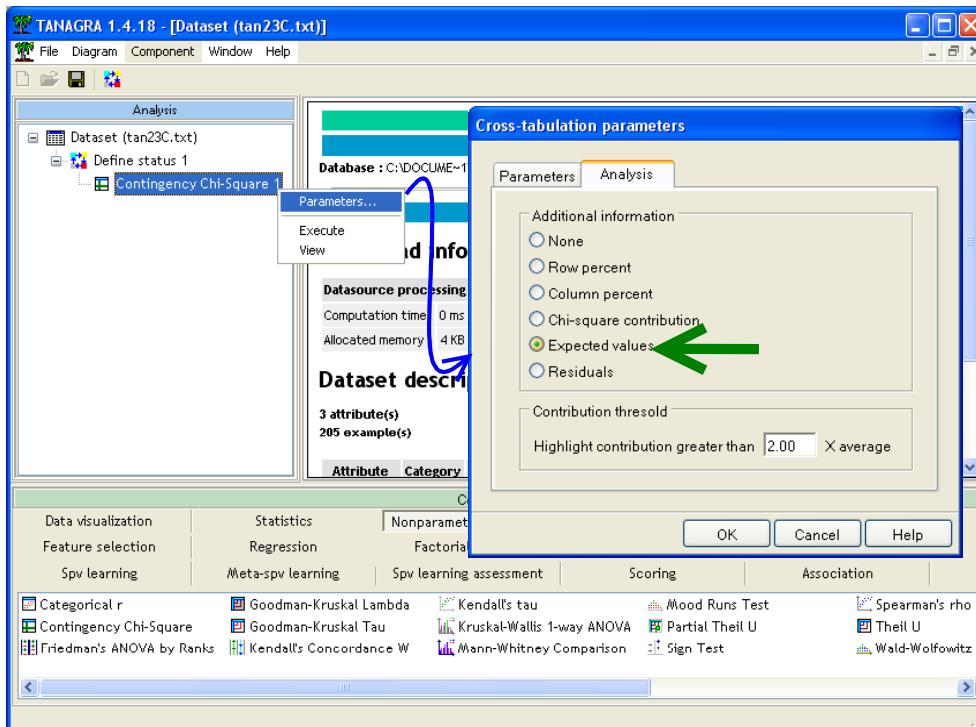
For comparison, here are the results of the STATISTICA software.



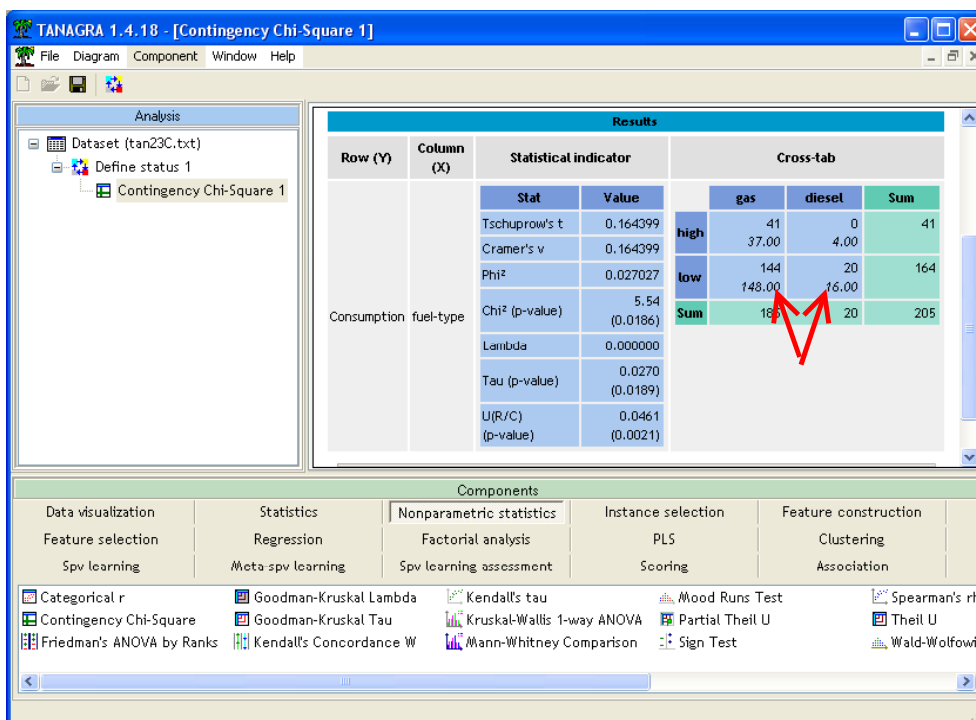
TANAGRA computes some additional statistical indicators derived from CHI-2, e.g. PHI-2, Tschuprow's t, Cramer's v.

Chi-square contributions and calculations details. It is possible to obtain the detail of calculations in the contingency table. We can for instance visualize the expected frequencies table:

we activate contextual menu PARAMETERS, and in ANALYSIS tab we click on the EXPECTED VALUES option.

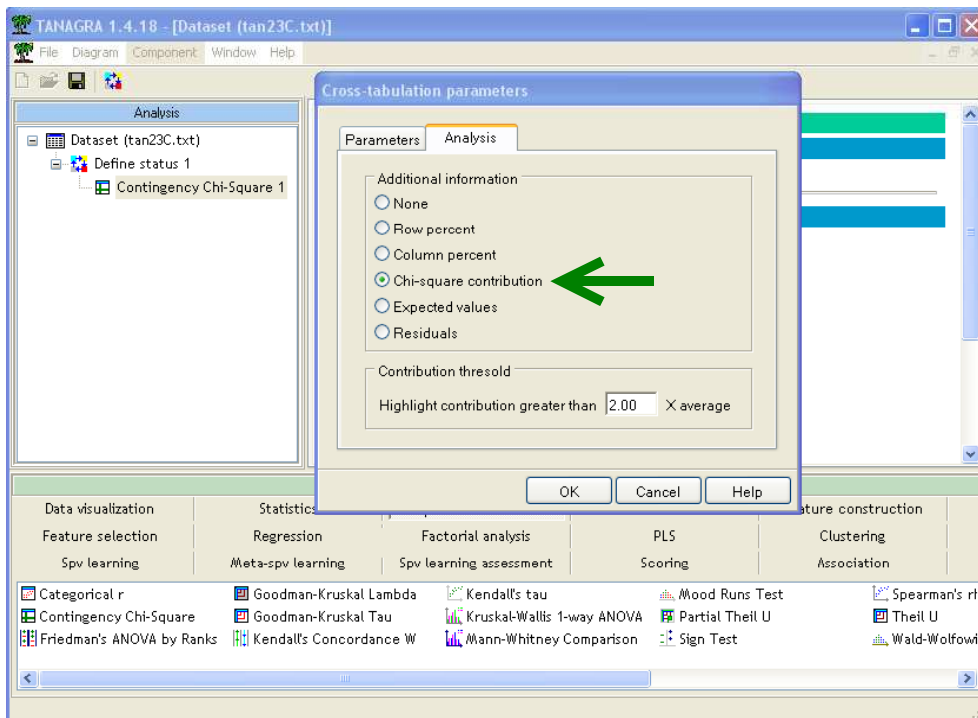


We click on the VIEW menu.

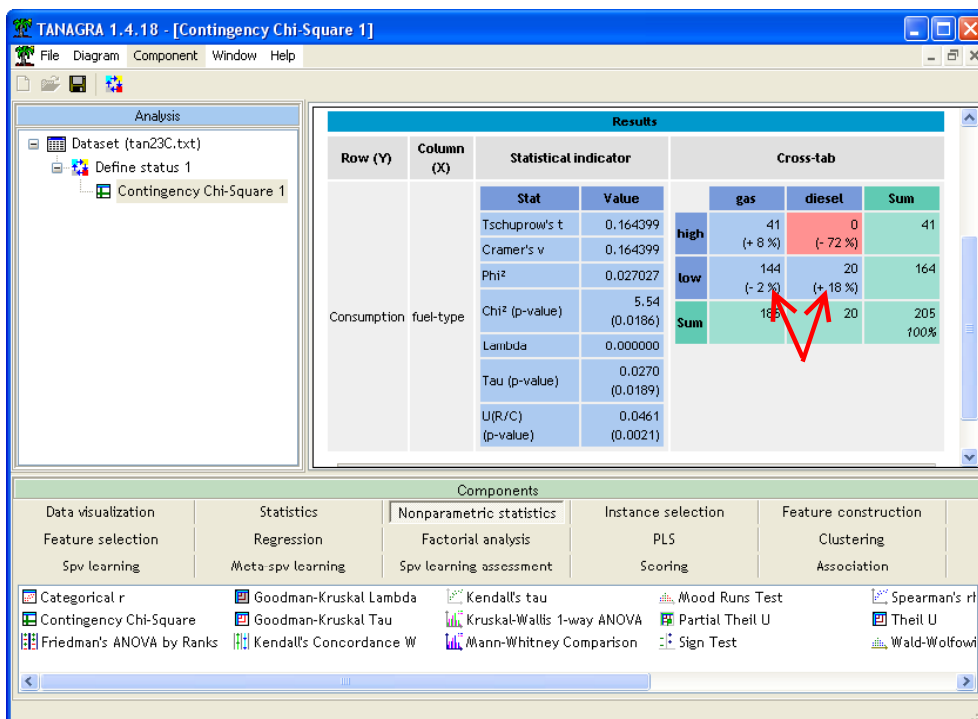


The expected frequencies are now printed under the observed frequencies.

We can also obtain the cells contributions. They are useful in determining which ranges of the two variables depicted in a table are contributing the most to the overall relationship. We click again on the PARAMETERS menu and select, in the ANALYSIS tab, the Chi-square contributions option.



We obtain the following results.



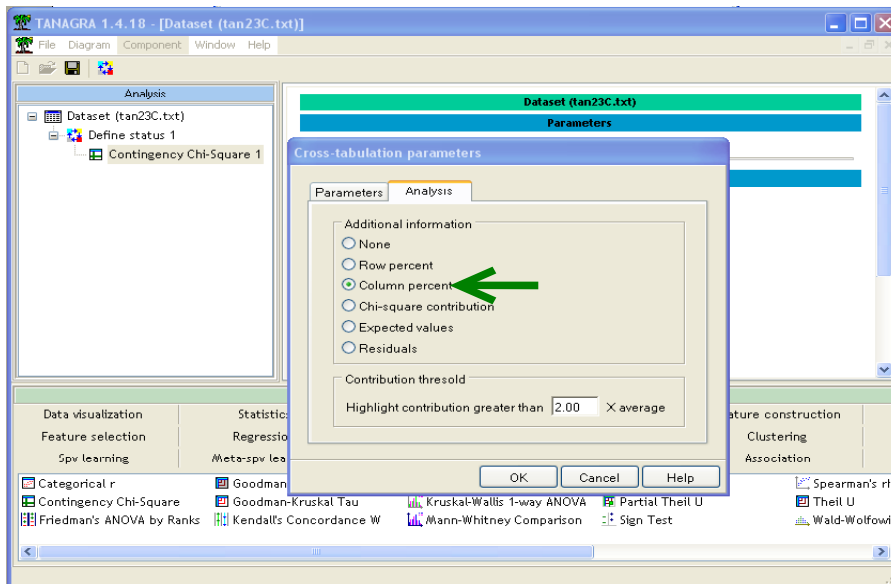
The contributions are displayed in percent of the whole chi-square statistic (CHI-2 = 5.54 in our dataset). When the contribution is 2 times (we can modify this) larger than the average contribution, it is underlined.

In our table, we observe that the relation between CONSUMPTION and FUEL-TYPE relies mainly on the opposition (-72%) between diesel fuel type and high consumption.

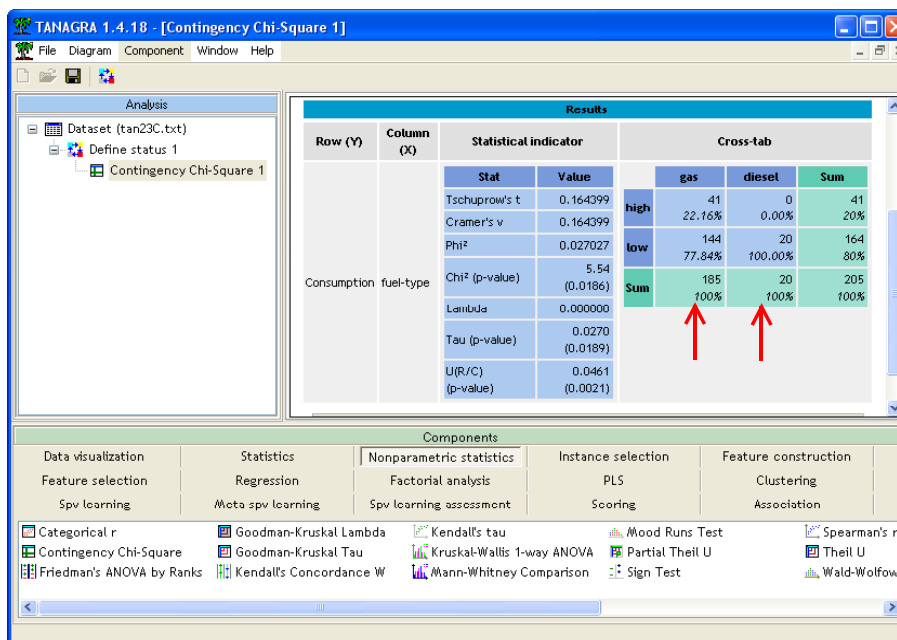
Of course, this kind of analysis is only interesting if the CHI-2 is significant.

Asymmetric association

In the most of cases, the variables have not a symmetrical role in the analysis. We want to check if one variable determines the values of another variable. For instance, we want to know if the FUEL-TYPE determines the CONSUMPTION. A first way to check this kind of association is to display the column percentage in the table. Then we can determine if the proportion of cars with low (high) consumption are different according to their fuel type. To do that, we modify again the parameters of the component.



We obtain the following table.



We note that the consumption is different according to whether the car uses diesel or gas. Working on percentages makes the columns comparable, which authorizes this kind of analysis.

PRE measures (Proportionate Reduction in Error)

PRE measures are asymmetric. They allow characterizing a prediction analysis (TARGET vs. INPUT). Three measures are available: Goodman & Kruskal's Lambda; Goodman & Kruskal's Tau; Theil's U (Uncertainty Coefficient). The last two indicators display also the p-value of the significance test.

It is possible to obtain the details of calculations using specific components.

Goodman-Kruskal Tau. We add the GOODMAN-KRUSKAL Tau component (NONPARAMETRIC tab) into the diagram, under DEFINE STATUS 1. We obtain the following results when clicking on the VIEW menu.

The screenshot shows the TANAGRA 1.4.18 software interface. The main window displays the results for the Goodman-Kruskal Tau component. The results are summarized in the following table:

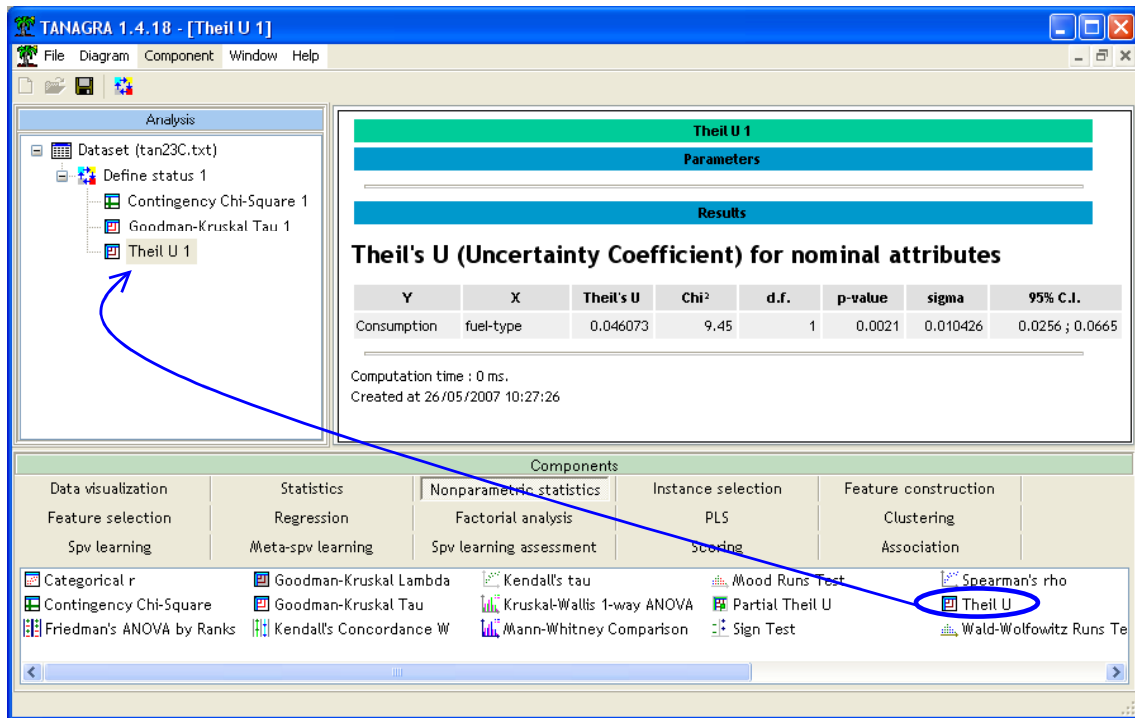
Y	X	Tau	Chi ²	d.f.	p-value
Consumption	fuel-type	0.027027	5.51	1	0.0189

Additional information displayed includes: Computation time : 0 ms. Created at 26/05/2007 10:23:51.

Tau statistic between CONSUMPTION (TARGET) and FUEL-TYPE (INPUT) is displayed again (0,0270).

The computed chi-square is a transformation of Tau (see Light & Margolin, 1971). It is 5.51. The degree of freedom is the same than the test of independence. The obtained p-value (0.0189) enables to conclude to a significant association at the 5%-level.

Theil's U. We add now the THEIL U component into the diagram.



The Theil's U is 0,046 in our analysis. There is also a chi-square transformation of this statistic, says also “Likelihood Ratio Chi-Square” statistic in some software. Then we can compute the *p-value* of the test. We observe the same conclusion as the preceding test.

For comparison, STATISTICA gives the following results.

STAT. ELEMENT.	Chi ²	dl	p
Chi ² de Pearson	5.540541	df=1	p= .01858
Chi ² du MV	9.452478	df=1	p= .00211
Coeff. d'incertitude	X= .0460726	Y= .0721161	X Y= .05622

For the Theil’s statistic, TANAGRA computes also the asymptotic variance/standard-deviation. Then, to give an idea of the variability of the solution, the 95% confidence interval is presented.

Partial association

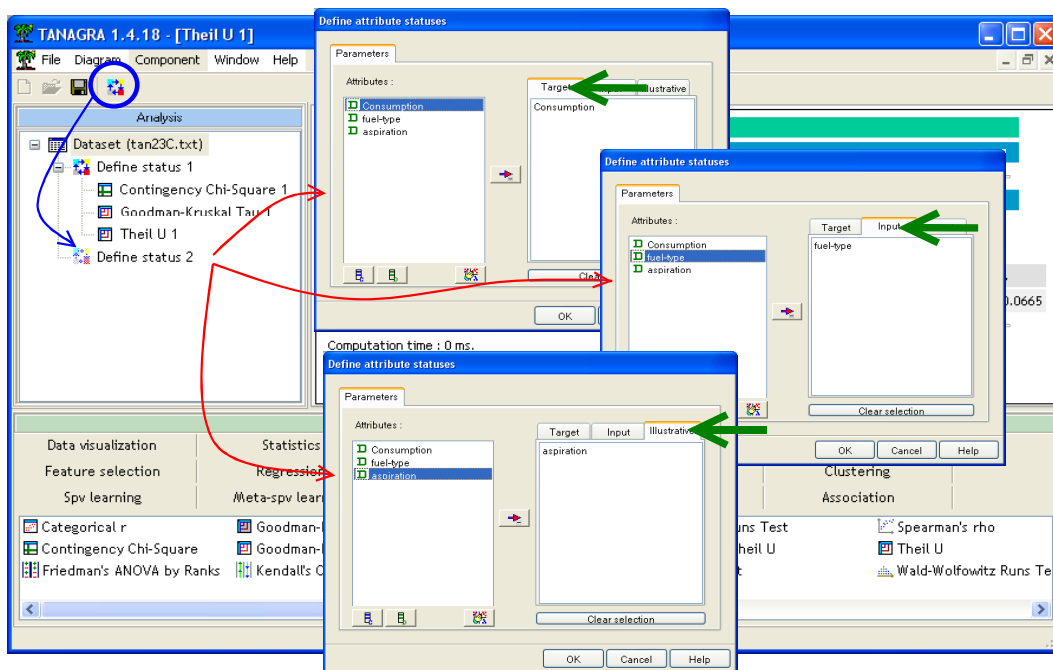
In some situations, we may seek to control for the effects of a variable E on the association between two nominal variables A and B.

For more details about this problem, see the following websites:

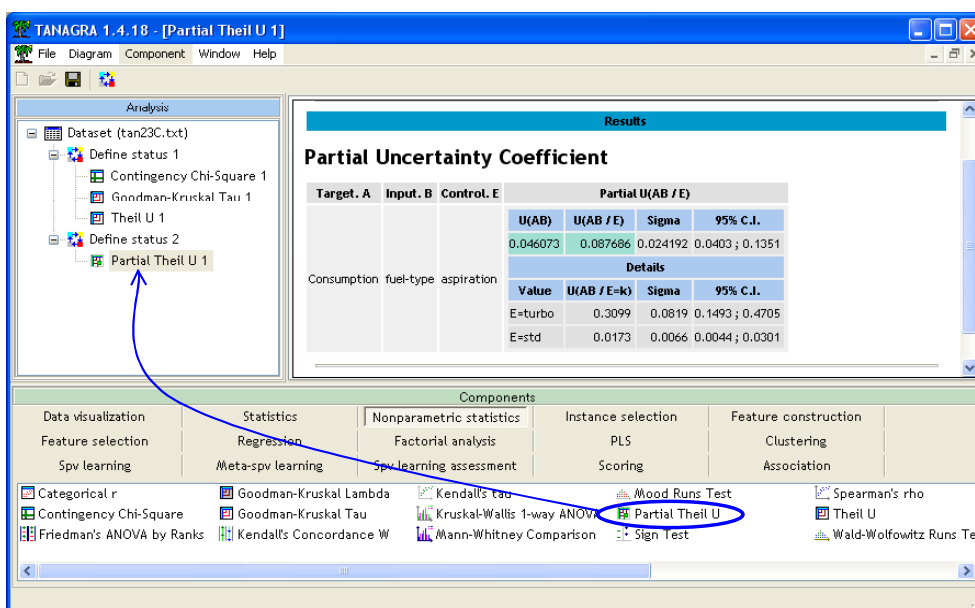
- <http://www2.chass.ncsu.edu/garson/PA765/association.htm#spss>
- <http://www2.chass.ncsu.edu/garson/pa765/partialr.htm>

In our dataset, ASPIRATION is the third variable that can play an important role (standard or turbo).

Partial Theil's U. We add a DEFINE STATUS component at the node of the diagram. We set CONSUMPTION as TARGET, FUEL-TYPE as INPUT, and ASPIRATION as ILLUSTRATIVE.



Then, we add the PARTIAL THEIL U component (NONPARAMETRIC STATISTICS tab).



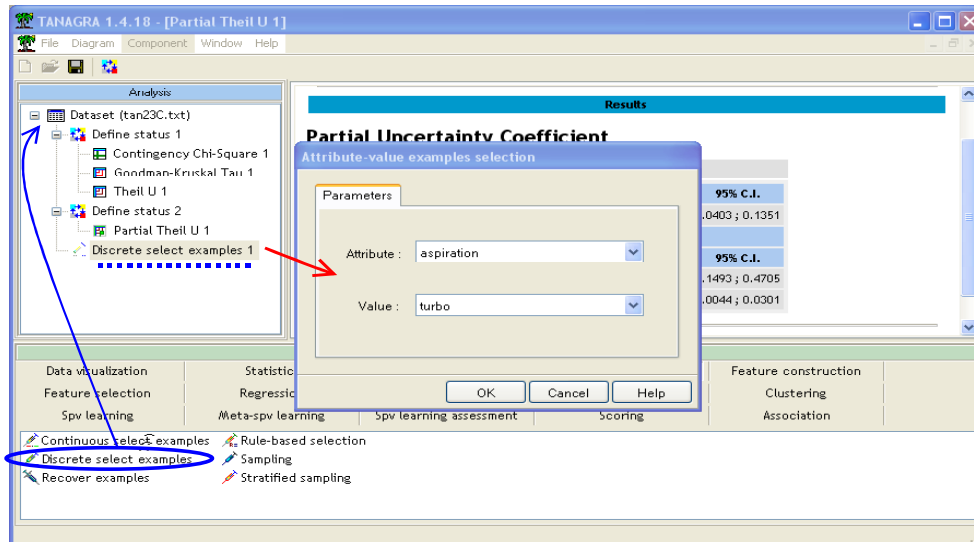
We observe that the partial association $U(AB/E) = 0.0876$ is larger than the association $U(AB) = 0.0460$. It seems that the association between CONSUMPTION and FULE-TYPE is more significant for ASPIRATION=TURBO cars. Indeed, we see in the table that the conditional association $U(AB/E=turbo)=0.3099$. TANAGRA computes also the asymptotic standard deviation and the 95% confidence interval.

The partial association is a weighted average of the conditional association index. The weight depends on the frequencies of the control value but also on the prediction error. See Olszak and Ritschard (1995) for detailed explanation².

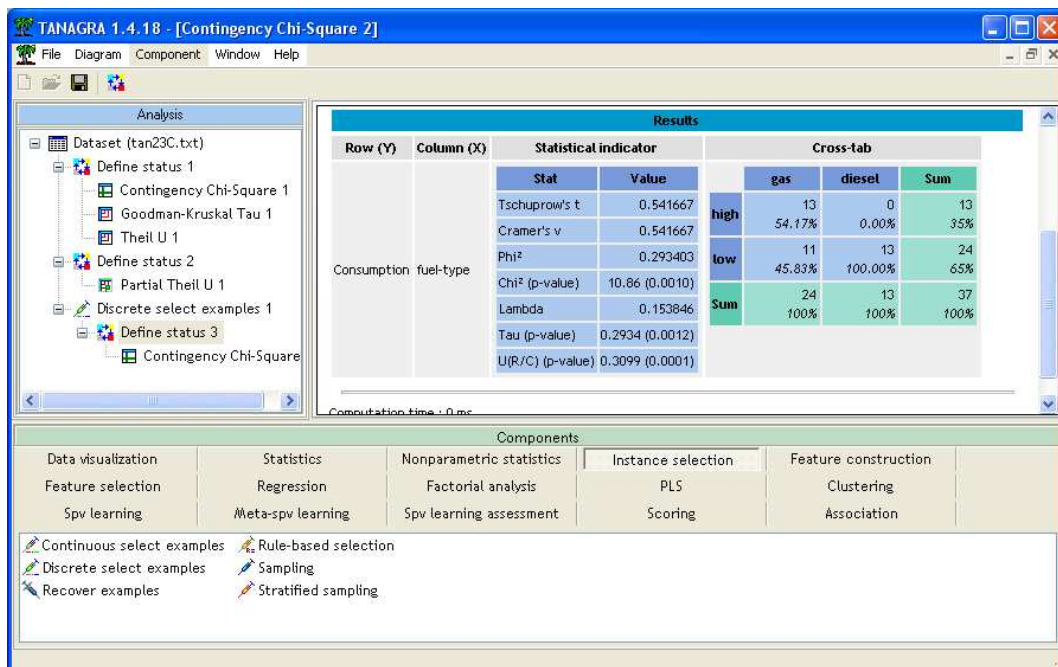
² M. OLSZAK, G. RITSCHARD (1995) « The behavior of nominal and ordinal partial association measure », in The Statistician, vol.44, n°2, pp.195-212.

Conditional association. We want to lead a deeper analysis on the subset of cars ASPIRATION=TURBO. We must apply all the computations on the examples corresponding to this characteristic.

To do that, we add the DISCRETE SELECT EXAMPLES component (INSTANCE SELECTION tab) into the root of the diagram. We set the attribute selection to ASPIRATION and the value selection to TURBO.



Then, we insert again the DEFINE STATUS and CONTINGENCY CHI-SQUARE components. TARGET attribute is CONSUMPTION, and the INPUT one is FUEL-TYPE. We obtain the following results when we click on the VIEW menu.

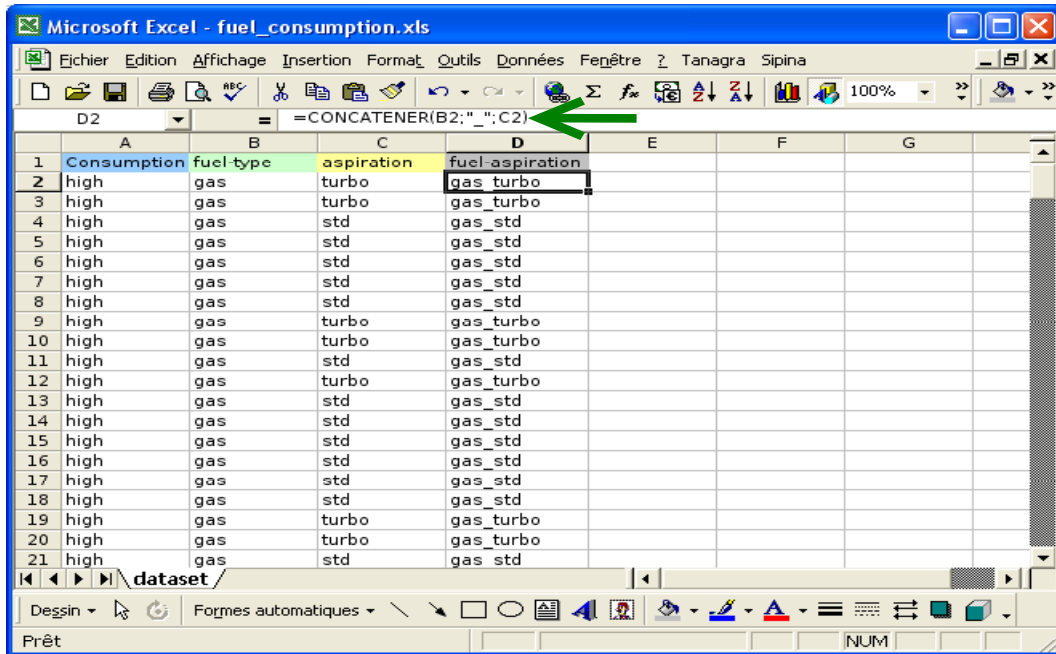


There are 37 ASPIRATION=TURBO cars in the database. All the FUEL-TYPE=DIESEL cars present a low CONSUMPTION. At the opposite, a majority (54.17%) of FUEL-TYPE=GAS cars have a high consumption. The association is significant at the 1% level (p-value = 0.0001).

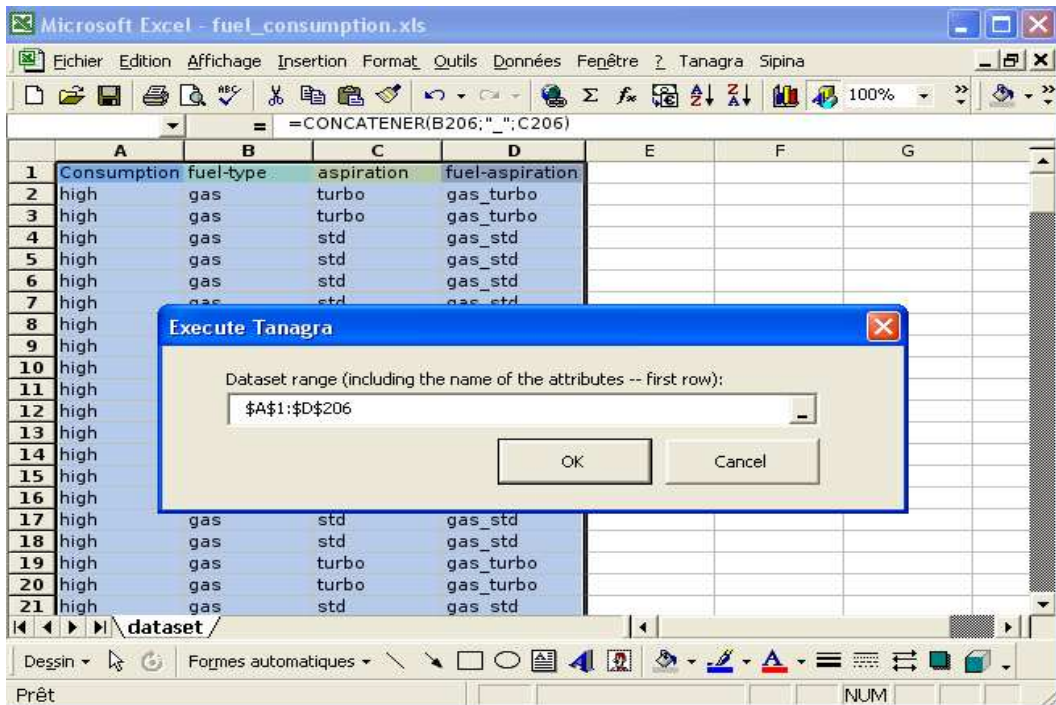
Variables interaction

Another kind of analysis is to analyze the interactions between FUEL-TYPE and ASPIRATION in the determination of CONSUMPTION. The easiest way is to create a new variable as follows.

Creation of a new variable. We close TANAGRA. In the EXCEL spreadsheet, we add a new column (FUEL-ASPIRATION). We insert the following formula³.

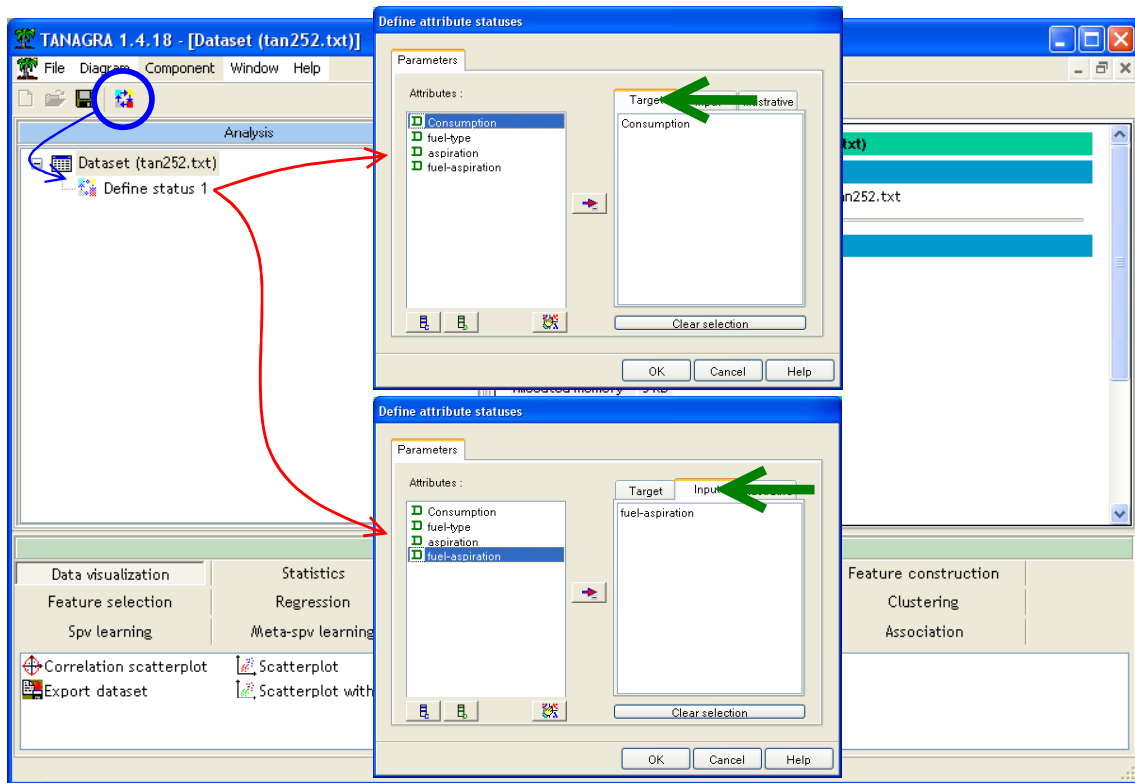


New diagram. We select the cells range and click on the TANAGRA/EXECUTE TANAGRA menu.

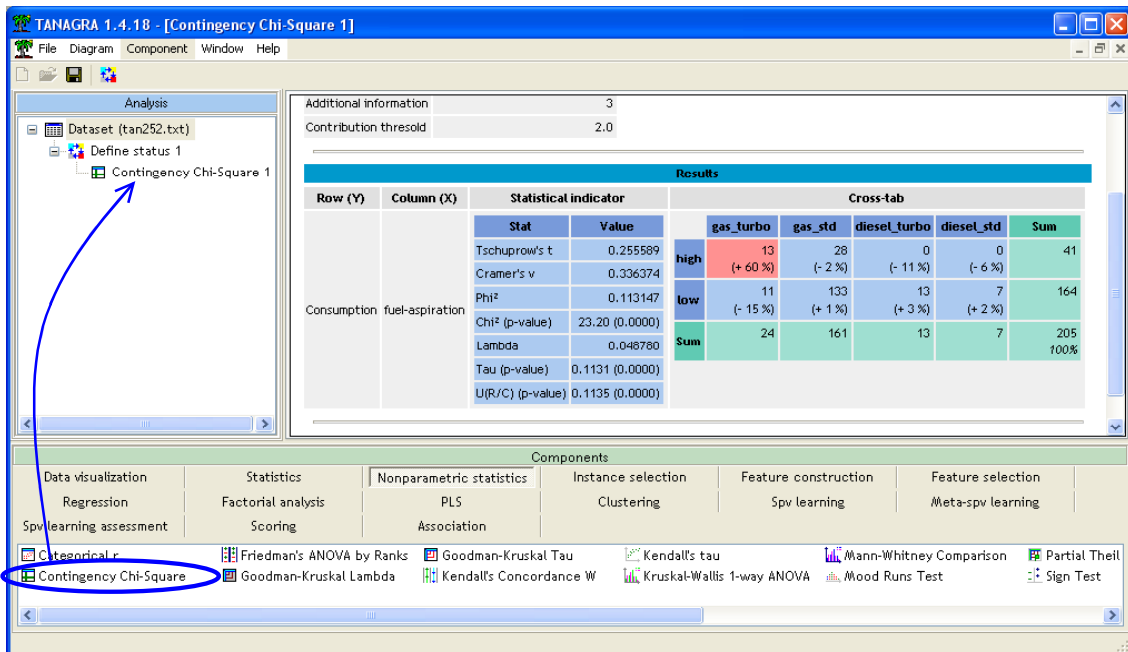


³ In English EXCEL version, we use the following function =CONCATENATE(...).

We insert the DEFINE STATUS component. We set CONSUMPTION as TARGET, FUEL-ASPIRATION as INPUT.



We add the CONTINGENCY CHI-SQUARE component. We select the Chi-Square contributions option.



The CHI-QUARE statistic is 23.20. The relation is very significant (p-value < 0.0000...). The main information (60%) into the contingency table is the "high" consumption of TURBO-GAS cars.

Conclusion

CHI-SQUARE statistic is widely used in order to gauge the relation between two nominal variables. In this tutorial, we have shown how to implement this kind of analysis with TANAGRA.

In addition, because the chi-square and derived measures are symmetric, they cannot handle a prediction/explanation process, we present also some asymmetric measures.