1 Topic

Non parametric tests for differences in location. (K ≥ 2) independent samples.

The tests for comparison of population try to determine if K (K ≥ 2) samples come from the same underlying population according to a dependent variable (X). In other words, we try to determine if the underlying distribution of X is the same whatever the group.

We talk about non parametric tests when we do not make assumption about the shape of the distribution of the dependent variable. They are considered as being "distribution free" methods, at the opposite of the parametric approaches.

In this tutorial, we implement various tests for differences in location. The Kruskal-Wallis test is certainly the most used one when we try to determine if the scores among groups are stochastically the same. But other tests exist. We compare the results obtained. We will complete the analysis by conducting multiple comparisons in order to identify groups that differ significantly from each other.

2 Dataset

The dataset comes from the Pr Richard Lowry's website[1]. One asked to n = 21 individuals to evaluate 3 kinds of wine (n_1 = 8 individuals belong to the first group; n_2 = 7 and n_3 = 6). Actually, the wine is exactly the same regardless of the group. The only difference is in the texture of the interview which was conducted in different ways (enthusiastic for the group A; less for group B; neutral for group C).

The dependent variable is RATING; its values vary from 1 (very bad) to 10 (very good). A further interesting aspect of this tutorial is to study the behavior of parametric methods on the same dataset (e.g. ANOVA, WELCH ANOVA[2]).

3 Kruskal-Wallis test

Our data file name is **wine_evaluation_nonparametric.xls**[3]. The individuals are described by two variables: RATING which is the dependent variable; GROUP which specifies the group membership.

3.1 Importing the dataset

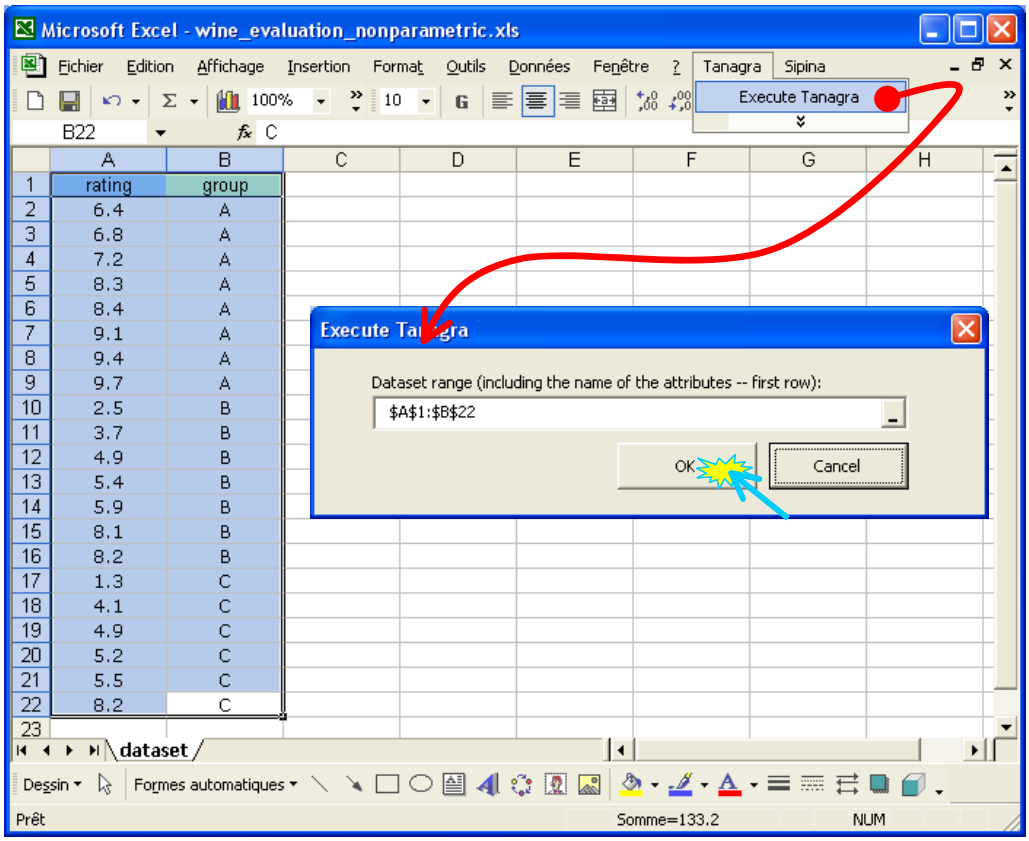
The simplest way to launch Tanagra is to open the data file into Excel. We select the data range; then we click on the Tanagra menu installed with the TANAGRA.XLA add-in[4]. After we checked the coordinates of the selected cells, we click on OK button.

[1] http://faculty.vassar.edu/lowry/ch14a.html ; the used formulas are described on the website.

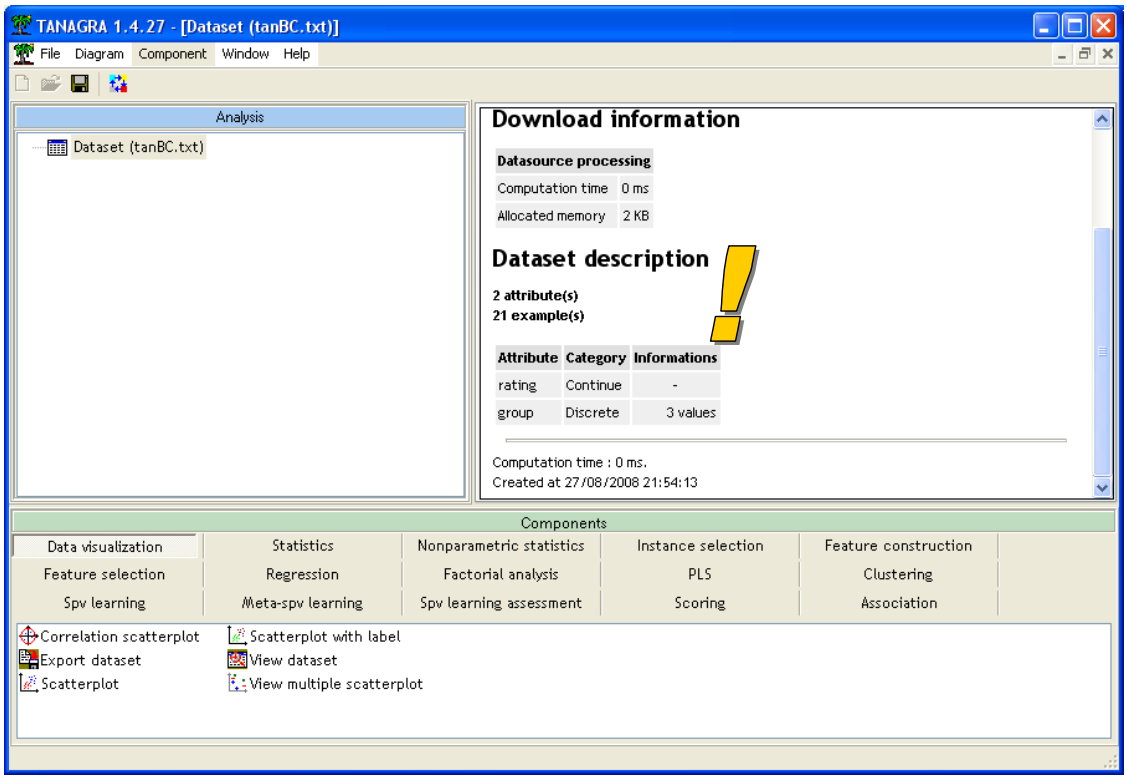
[2] See http://data-mining-tutorials.blogspot.com/2009/11/parametric-tests-for-comparing.html

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

[4] See http://data-mining-tutorials.blogspot.com/2008/10/excel-file-handling-using-add-in.html

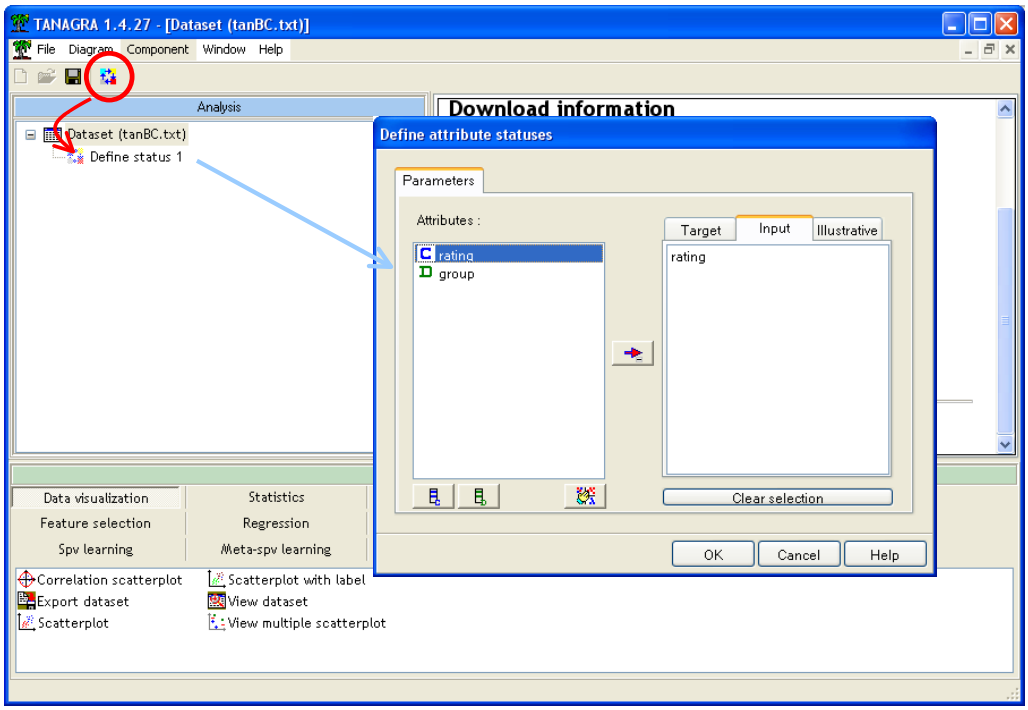


TANAGRA is automatically launched. A new diagram is created and the dataset is imported. We have 210 instances and 2 variables.

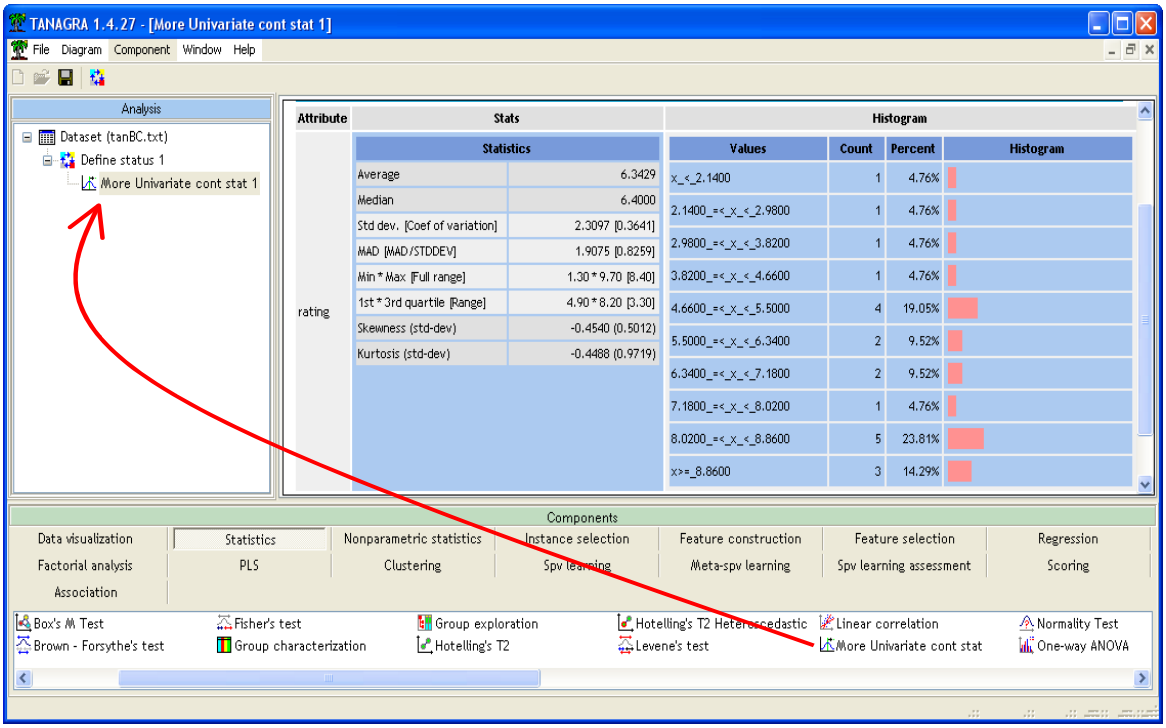


3.2 Descriptive statistics

In a first step, we want to describe our dataset by computing some descriptive statistics indicators on the dependent variable. We add the DEFINE STATUS component into the diagram. We set RATING as INPUT.

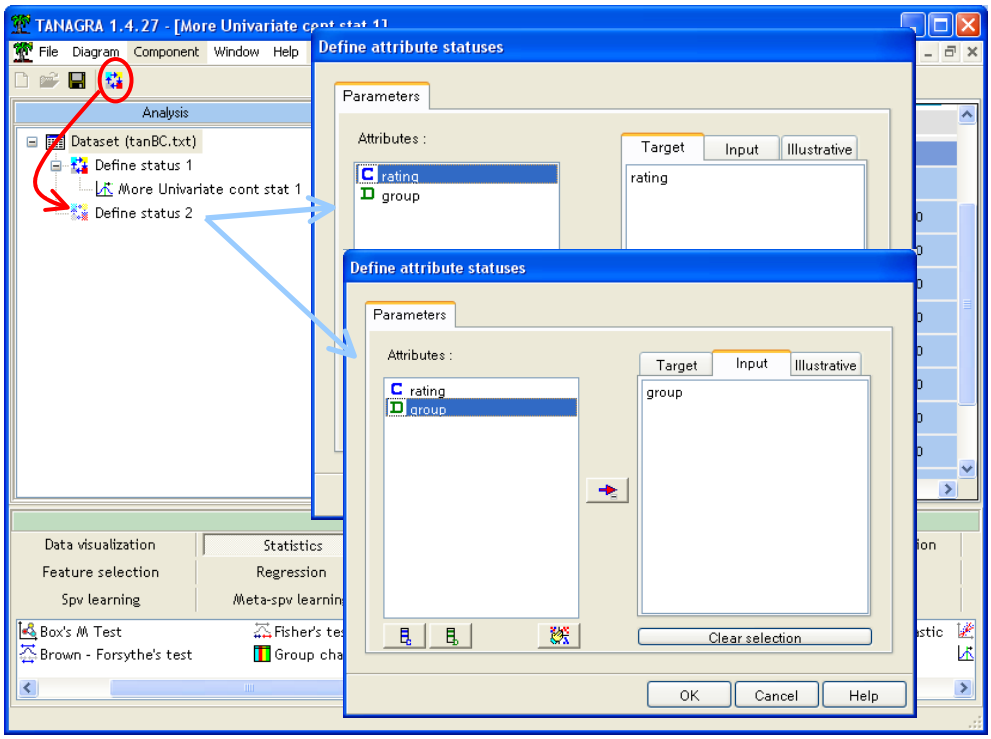


Then we insert the MORE UNIVARIATE STAT component (STATISTICS tab). We click on the VIEW menu in order to obtain the following results.

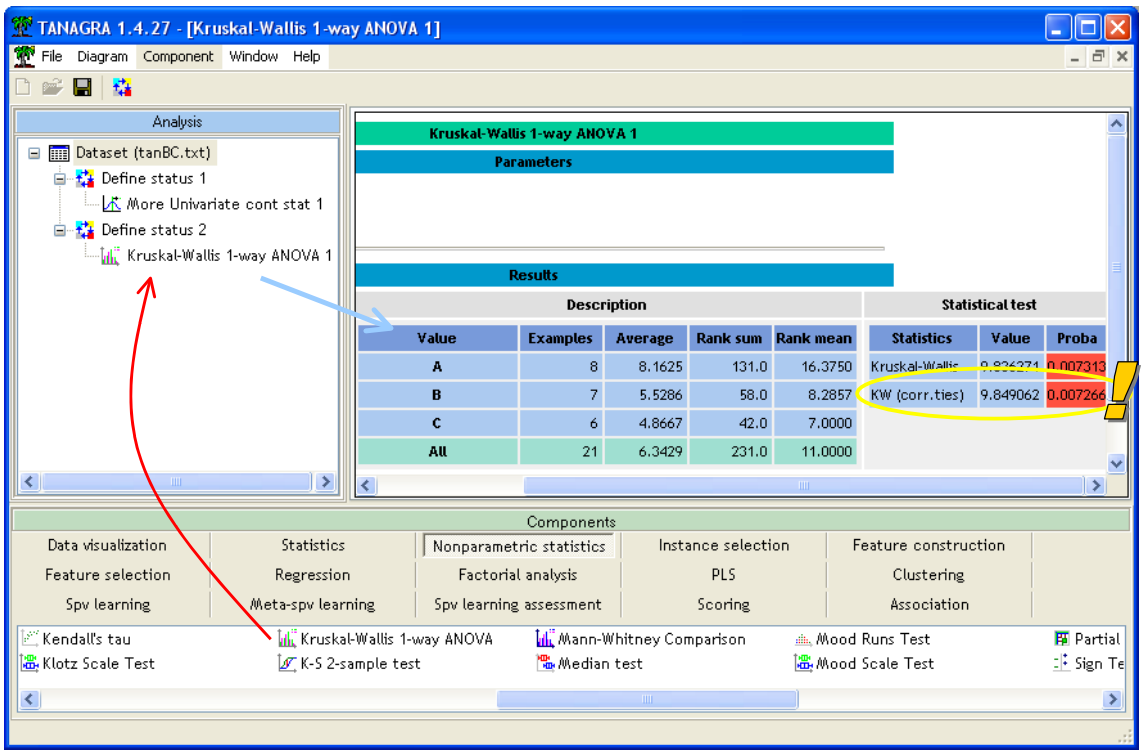


3.3 Kruskal-Wallis test

We want to check the difference in rating among groups. We insert the DEFINE STATUS component into the diagram. We set RATING as TARGET, GROUP as INPUT.



We add the KRUSKAL-WALLIS 1-WAY ANOVA component (NONPARAMETRIC STATISTICS tab). We click on the VIEW menu.



TANAGRA computes the test statistic with and without correction for ties. For this last one, we note that the test statistic is KW = 9.849062. Under the null hypothesis, it follows a chi-square distribution with (K – 1 = 3 – 1 = 2) degrees of freedom. The p-value of the test is p = 0.007266. We conclude that the rating is different for at least one group at the 5% significance level.

3.4 Detecting the origin of the deviation

When the null hypothesis is rejected, the next step is the detection of the groups which differ significantly to the others. We perform a multiple comparison between groups. We conclude that the deviation between two groups (i and j) is significant if we observe the following situation[5].

$$\left|\bar{r}_i - \bar{r}_j\right| \geq u_{1-a}\sqrt{\frac{n(n+1)}{12}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

Where

- n is the size of the overall dataset;
- n_i (resp. n_j) is the size of the group i (resp. j);
- $\bar{r}_i$ (resp. $\bar{r}_j$) is the mean of the rank for the group i (resp. j) ;
- α is the overall significance level for the Kruskal-Wallis test (5% in our context);
- $a = \frac{\alpha}{K(K-1)}$ is the significance level for the individual test (comparison);
- u_{1-a} is the (1-a)-th quantiles for the standard normal distribution.

We give the detail of computation for the comparison of groups A and B:

- $n_A = 8$ et $n_B = 7$;
- $\bar{r}_A = 16.375$, $\bar{r}_B = 8.2857$, $\left|\bar{r}_A - \bar{r}_B\right| = \left|16.375 - 8.2857\right| = 8.089$;
- $\alpha = 0.05 \rightarrow a = \frac{0.05}{3(3-1)} = 0.008 \rightarrow u_{1-0.0083} = 2.3940$
- $u_{1-a}\sqrt{\frac{n(n+1)}{12}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} = 2.3940\sqrt{\frac{21(21+1)}{12}\left(\frac{1}{8}+\frac{1}{7}\right)} = 7.688$

Since $\left|\bar{r}_i - \bar{r}_j\right| = 8.089 > 7.688$, the rating is significantly different according to the groups A and B.

We led the comparisons between all the groups, we observe that the deviation is statistically significant between A vs. B and A vs. C.

[5] S. Siegel, J. Castellan, « Nonparametric Statistics for the Behavioral Sciences », McGraw-Hill, Inc., 1988; pp.213-214.

Description					Statistical test		
Value	Examples	Average	Rank sum	Rank mean	Statistics	Value	Proba
A	8	8.1625	131	16.375	Kruskal-Wallis	9.836271	0.00731
B	7	5.5286	58	8.2857	KW (corr.ties)	9.849062	0.00727
C	6	4.8667	42	7			
All	21	6.3429	231	11			

n	21
K	3

alpha	0.05
a	0.0083

u	2.3940

Group.1	Group.2	|Difference|	Crit.Level	Significant
A	B	8.089	7.688	yes
A	C	9.375	8.022	yes
B	C	1.286	8.264	no

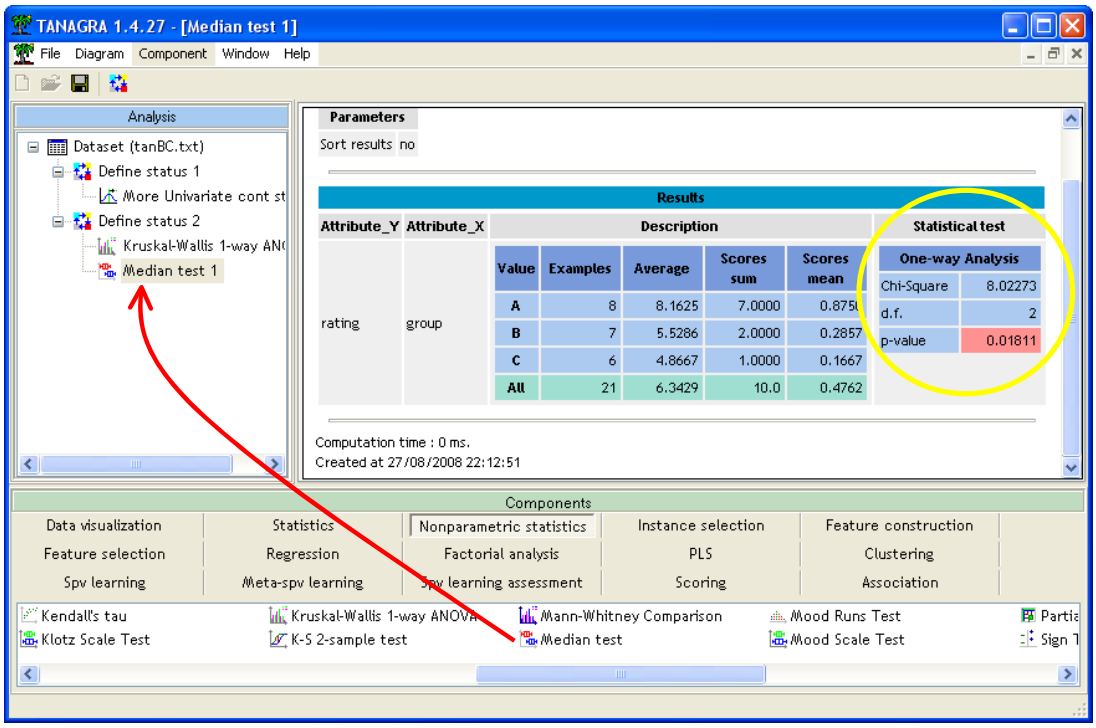
4 Other tests

The Kruskal-Wallis test is the most popular one when we want to compare differences in location among K groups with a non parametric test. But this is not the only one. Other techniques exist, little known and / or not implemented in tools. We present some of them in this section.

4.1 The median test

The median test for independent samples is an alternative of the Kruskal Wallis test. But it is in general less efficient. However, it is powerful for distributions that are symmetric and heavy-tailed[6].

We insert the MEDIAN TEST component (NONPARAMETRIC STATISTICS tab) into the diagram.



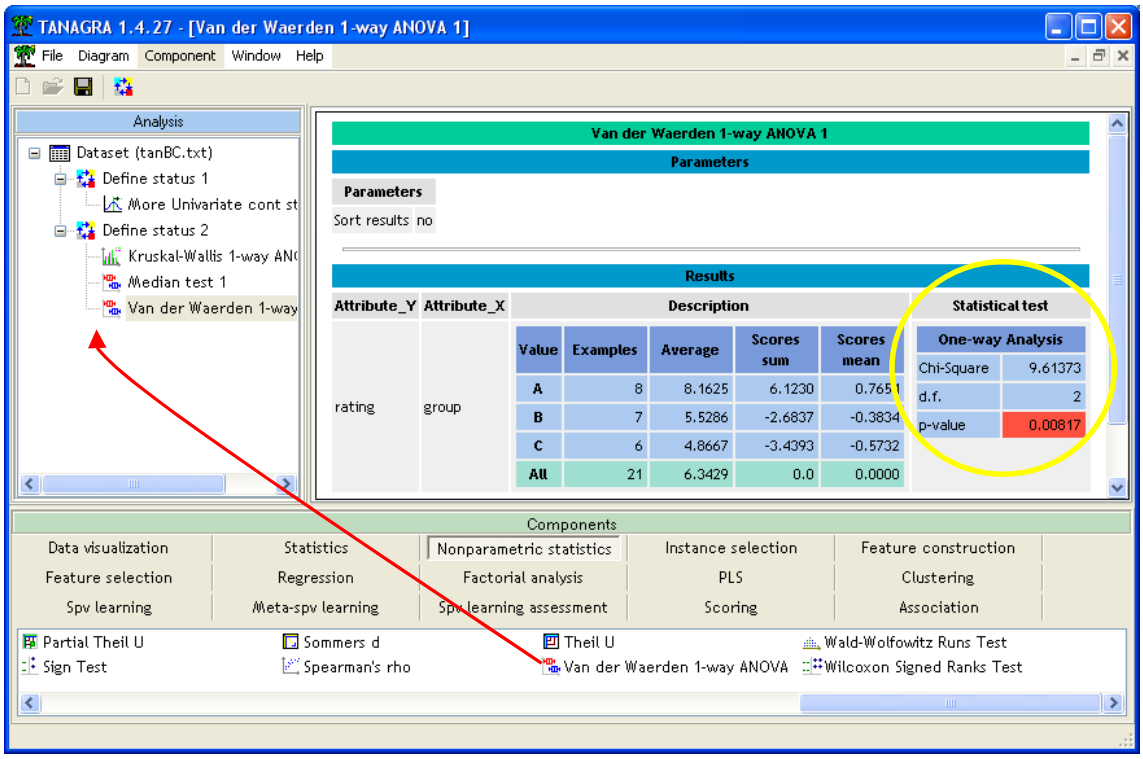
[6] See http://v8doc.sas.com/sashtml/stat/chap47/sect17.htm

The test statistic is $\chi^2 = 8.02273$; the p-value is $p = 0.01811$. We can reject the null hypothesis at the 5% significance level.

4.2 The Van der Waerden test

The Van der Waerden test uses the Van der Waerden scores, which convert the ranks to the quantiles of the standard normal distribution[7]. This test is especially suitable when the underlying distribution of the dependent variable is close to the normal distribution.

We insert the VAN DER WAERDEN 1-WAY ANOVA component into the diagram.



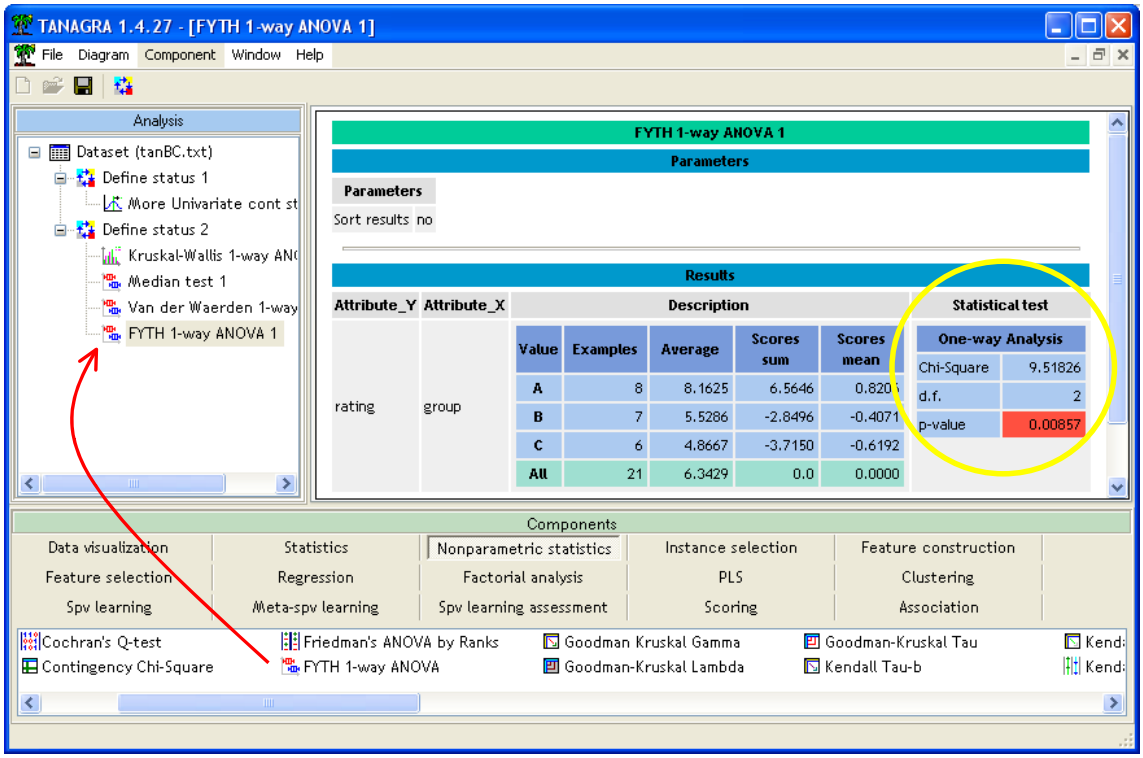
The conclusion is very similar to the Kruskal-Wallis test. This is not surprising given the characteristics of these methods and our data.

4.3 The Fisher-Yates-Terry-Hoeffding test

The Fisher-Yates-Terry-Hoeffding variant is very similar to the Van der Waerden test. It is based also on normal scores. But it uses the expected normal scores.

We add the FYTH 1-WAY ANOVA into the diagram. We obtain the following results when we click on the VIEW menu. The results are not surprisingly very similar to the Van der Waerden test.

[7] http://en.wikipedia.org/wiki/Van_der_Waerden_test



5 Conclusion

The majority of the components (MEDIAN TEST, VAN DER WAERDEN 1-WAY ANOVA and FYTH 1-WAY ANOVA) presented in this tutorial can handle a comparison of 2 independent samples. An additional result, the z-value variant of the chi-square statistic, is supplied in this context.

About the Kruskal-Wallis test, it is more suitable to use the MANN-WHITNEY COMPARISON component when we want to compare 2 independent samples.