

1 Subject

Nonparametric tests, 2 independent samples, differences in location.

The aim of homogeneity test (or test for difference between groups) is to check if K ($K \geq 2$) samples are drawn from the same population according to a variable of interest. In another words, we check if the probability distribution is the same in each sample.

The nonparametric tests make no assumptions about the distribution of the data. They are called also "distribution free" tests.

In this tutorial, we show how to implement nonparametric homogeneity tests for differences in location for $K = 2$ populations i.e. the distributions of the populations are the same excepting a shift in location (central tendency). For some of them, the test is more general. It checks if the data values in one group are stochastically larger (or smaller) than in the other.

The **Kolmogorov-Smirnov** test is the more general one. It checks all kind of differences between the cumulative distribution functions (CDF). Afterwards, we can implement other tests which characterize more deeply the difference. The **Wilcoxon-Mann-Whitney** test is certainly the most popular one. We will see in this tutorial that other tests can be also implemented.

Some the tests introduced here are usable when the number of groups is upper than 2 ($K > 2$).

2 Dataset

The dataset comes from the on-line course «STAT 500 – Applied Statistics»¹ (Penn State University). We are mainly concerned by the Lesson 10 which deals with the comparison of two-population means. We deal with the following problem: "In a packing plant, a machine packs cartons with jars. It is supposed that a new machine will pack faster on the average than the machine currently used. To test that hypothesis, the times it takes each machine to pack ten cartons are recorded".

The data seems compatible with a Gaussian distribution. If we use the usual parametric two-sample t-test, we obtain $t = -3.4$. The deviation is significant with a p-value $p = 0.0032$ for a two-tailed test. An interesting aspect of this tutorial is to study the behavior of the nonparametric tests on these data, and compare the results with those of the t-test.

3 Creating a diagram and importing the data file

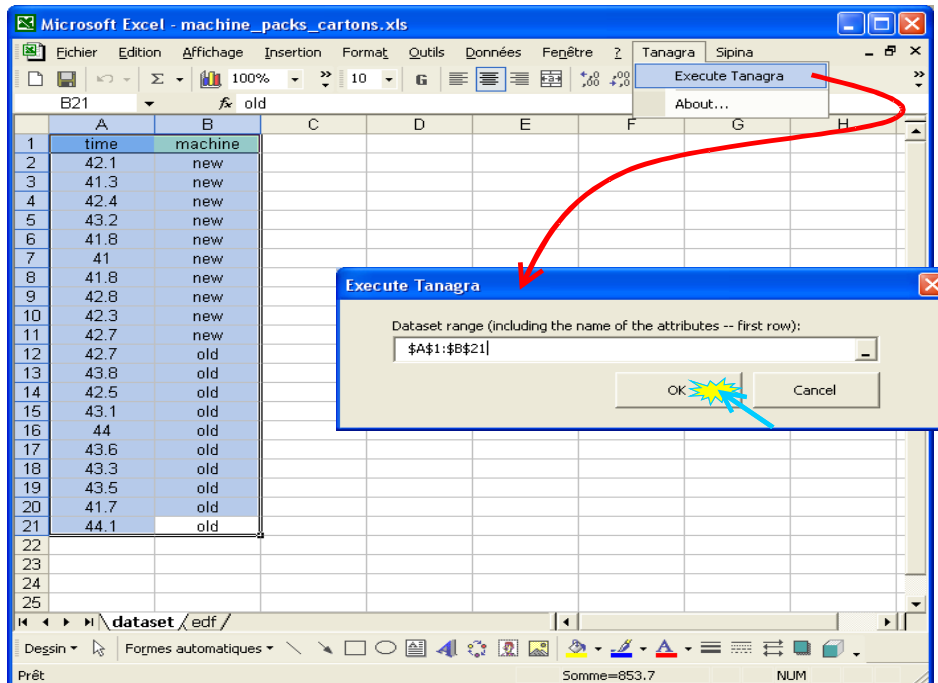
The data file is **machine_packs_cartons.xls**². There are 2 columns: the time to pack cartons (TIME, in seconds); and the types of the machine (MACHINE; « new » or « old »).

¹ http://www.stat.psu.edu/online/development/stat500_spss/index.html ; Lesson 10.

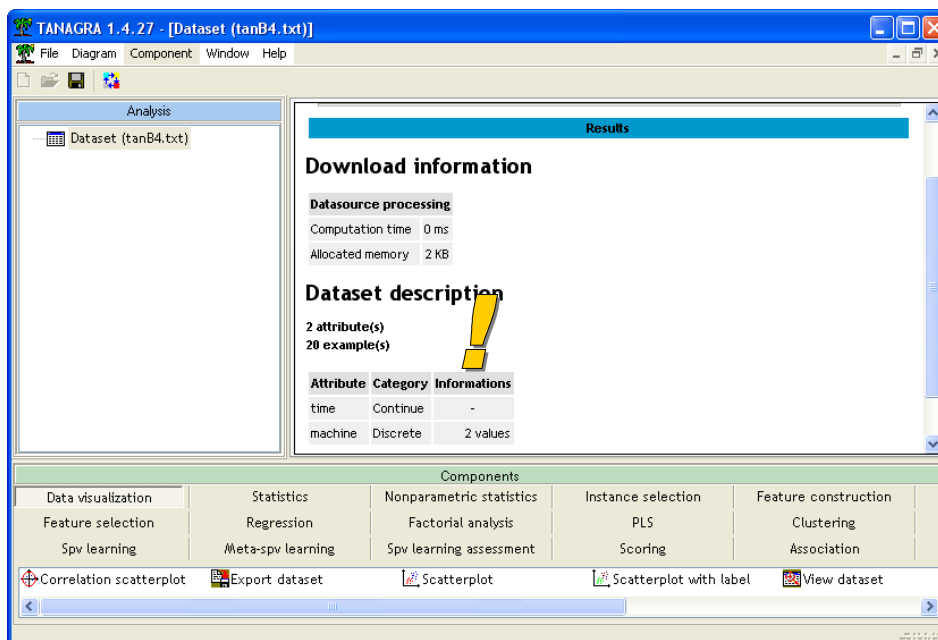
² http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/machine_packs_cartons.xls

3.1 Importing the data file

The easiest way to handle the dataset is to open the file into EXCEL spreadsheet. We select the range of cells then we click on the TANAGRA / EXECUTE TANAGRA menu which is installed with and add-in³. We check the selection. Then, we click on OK button.



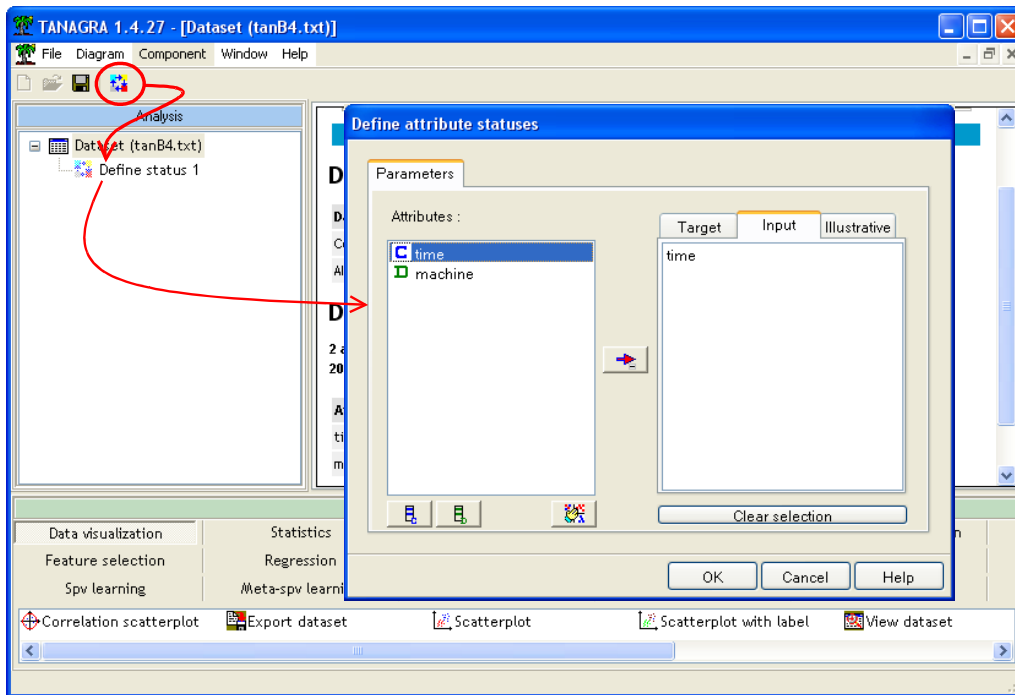
TANAGRA is automatically launched. A new diagram is created and the data imported. There are 20 instances and 2 variables.



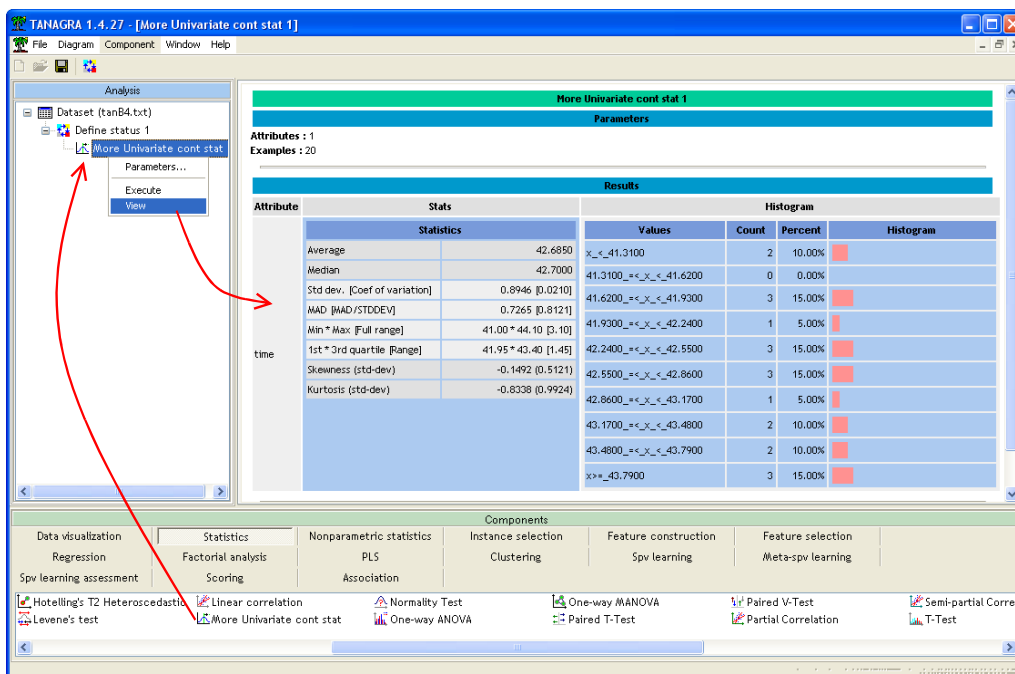
³ See <http://data-mining-tutorials.blogspot.com/2008/10/excel-file-handling-using-add-in.html> for details.

3.2 Descriptive statistics

We compute some descriptive statistics indicators, mainly for detecting unusual values. We add the DEFINE STATUS component into the diagram. We set TIME column as INPUT.



Then, we insert the MORE UNIVARIATE STAT component (STATISTICS tab). We obtain the following result when we click on the VIEW menu.



There are no particular comments here. We note only that the distribution is fairly uniform. Anyway, choosing automatically 10 intervals for our histogram seems not relevant for our dataset.

4 Comparing the cumulative distribution functions (CDF)

The kind of tests that we use in this section rely on the comparison of the conditional CDF i.e. the CDF of the variable in each group. In the following figure (Figure 1), we see a deviation between the CDF. If we analyze the median for instance, for $F(X)=0.5$, the median in the first group is 42.1, it is 43.3 in the second group.

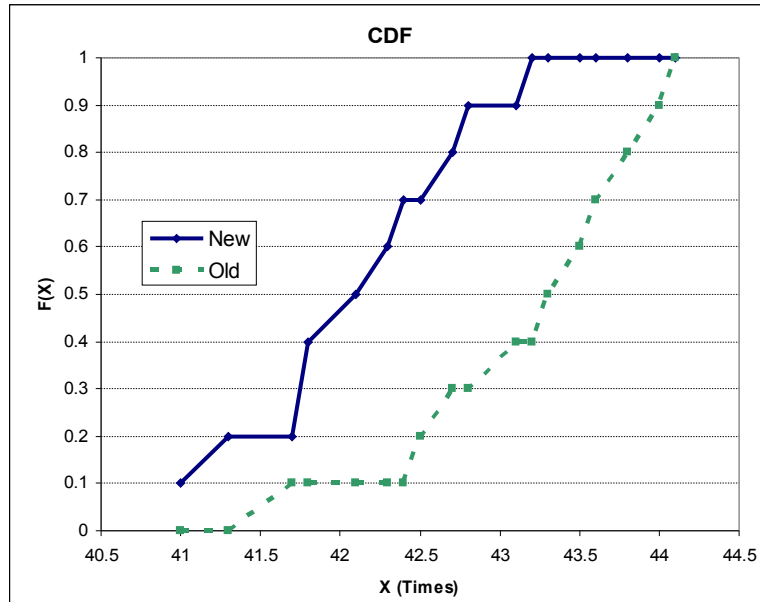
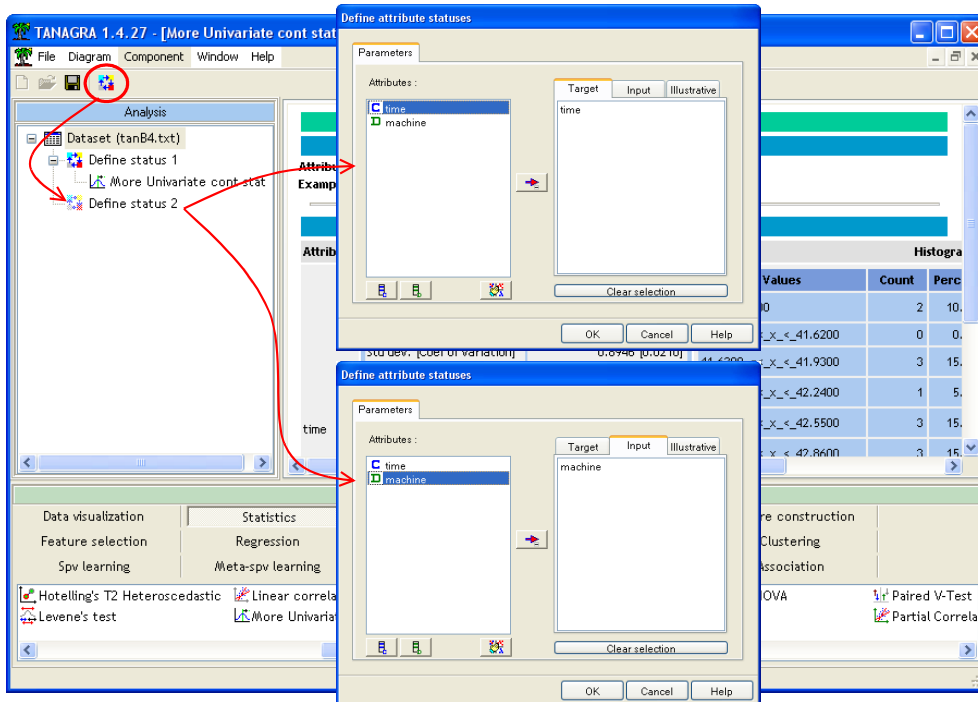
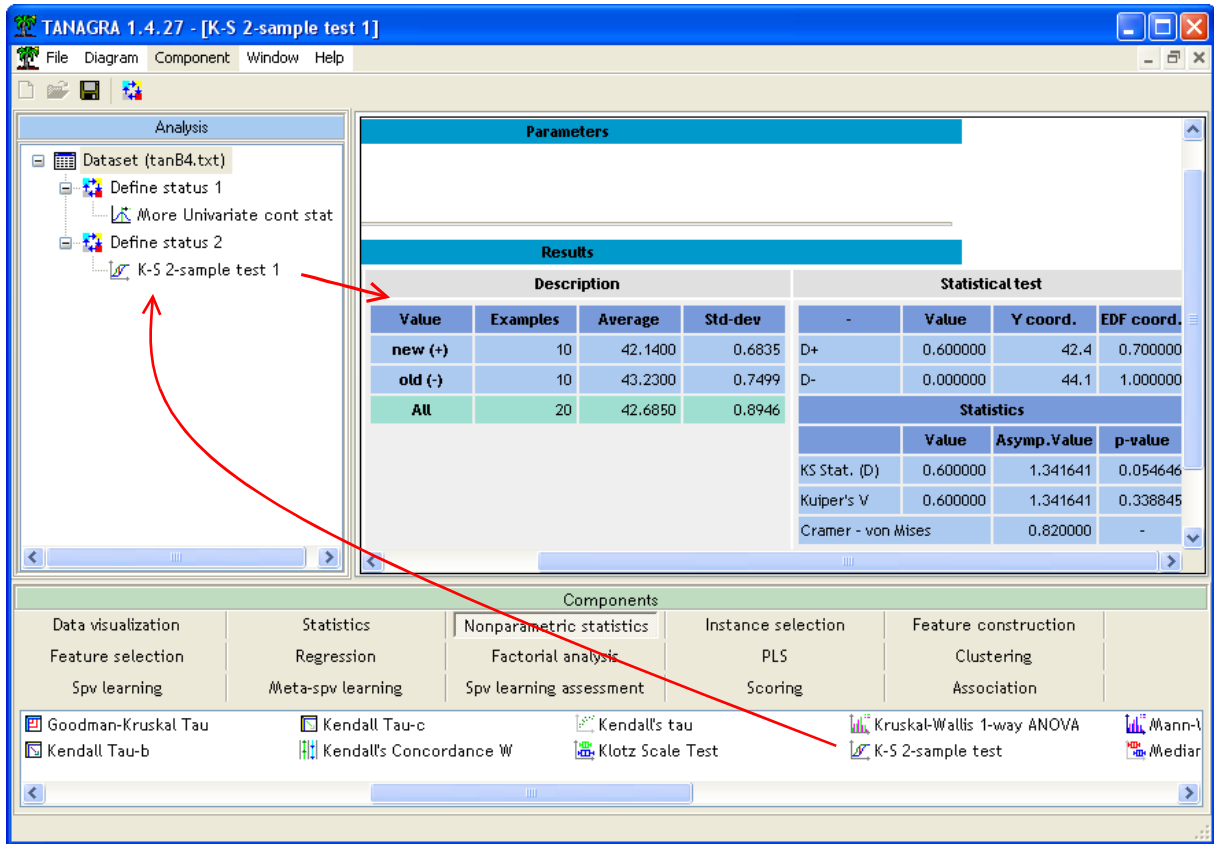


Figure 1 - Comparing the CDF according to the groups

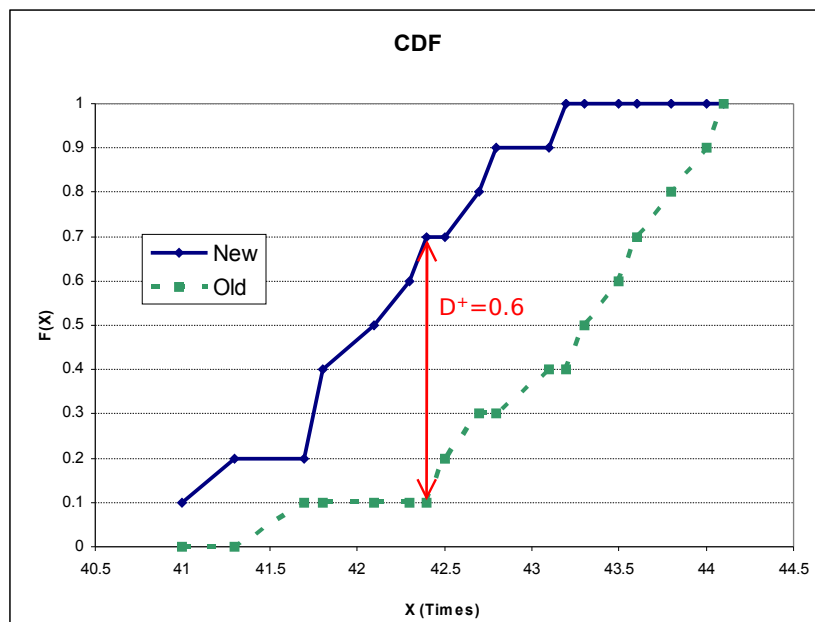
We try to validate this visual feeling by using the Kolmogorov-Smirnov test. We add the DEFINE STATUS component into the diagram, we set TIME, the dependent variable, as TARGET; MACHINE, group membership indicator, as INPUT.



Then we add the K-S 2-SAMPLE TEST (NONPARAMETRIC STATISTICS tab) component. We click on the VIEW menu. We obtain the following results.



The "positive" group is "MACHINE = NEW". The choice is arbitrary. The difference between the CDF is never negative ($D^- = 0$) i.e. the CDF of "OLD" is never above of the one of "NEW". The positive difference is maximum ($D^+ = 0.6$) when $X = 42.4$, with $F_+(X) = 0.7$.



The Kolmogorov-Smirnov statistic is $D = \max(D+ ; D-) = 0.6$. It is used to check the significance of the difference. By using the asymptotic distribution, with $d = \sqrt{\frac{10 \times 10}{10 + 10}} \times D = 1.341641$, we obtain the p-value $p = 0.054646$. At the significance level 5%, we conclude that the difference is not significant. The dataset is compatible with the null hypothesis i.e. the CDF are the same.

We must however put into perspective this result. We know that the KS test is an omnibus test. It tries to detect all kind of deviation. Thus, it is not very powerful with a high type II error.

In the following, we try to make deeper the analyze by characterizing the nature of the deviation.

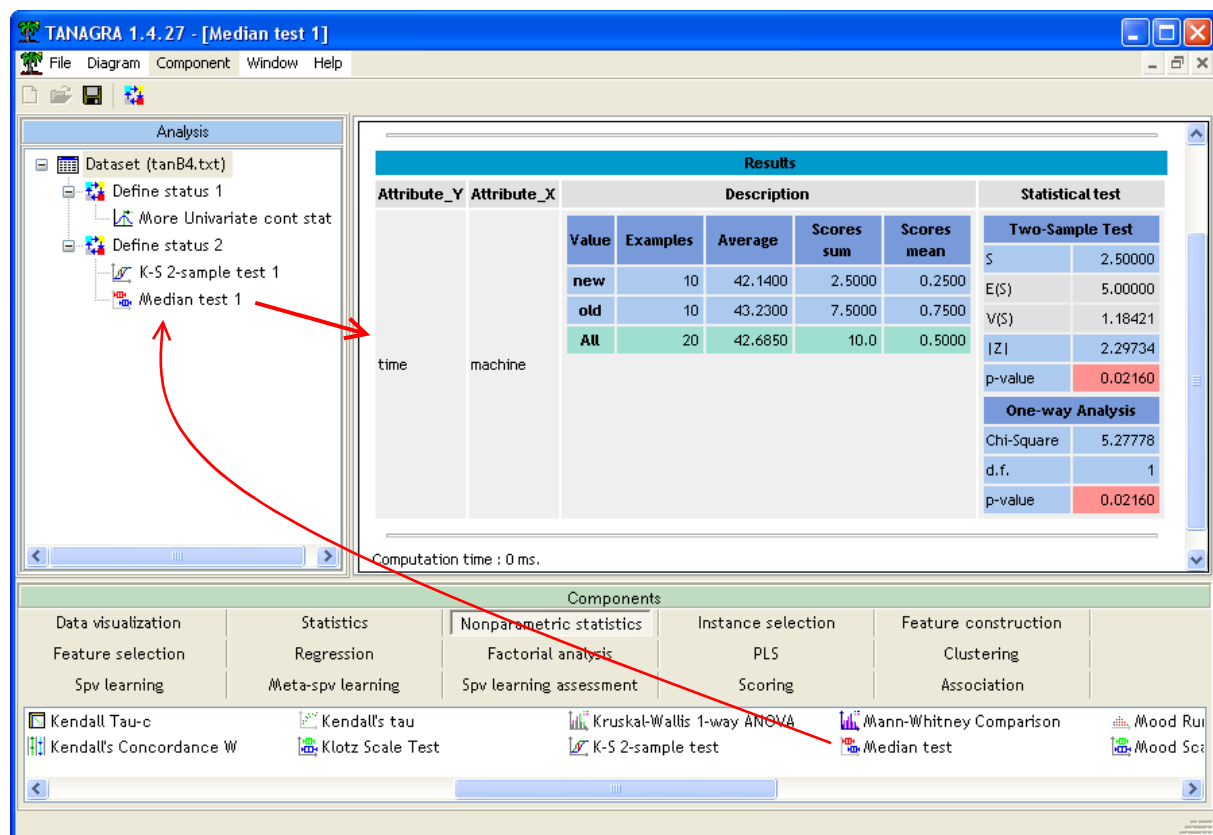
5 Comparison based on the location model

We must use a more specific test in order to detect the deviation between the groups. Here, we try to characterize the deviation in location (the central tendency) between the CDF.

5.1 Median test

This test compares the median of the distributions of each group. We have suggested this idea above, we use now a more rigorous approach based on a statistical test.

We add the MEDIAN TEST component (NONPARAMETRIC STATISTICS tab) into the diagram. We obtain the following results.



The test statistic is $S = S_1 = 2.5$. It corresponds to the sum of scores associated to the first group (MACHINE = NEW). For the second group, we obtain $S_2 = 7.5$.

In order to check the significance of the deviation, because under the null hypothesis the distribution of S is Gaussian, with $E(S) = 5.0$ and $V(S) = 1.18421$, the standardized statistic is

$$|Z| = \frac{|S - E(S)|}{\sqrt{V(S)}} = \frac{|2.5 - 5.0|}{\sqrt{1.18421}} = 2.29734$$

The p-value is $p = 0.0216$. Unlike the Kolmogorov-Smirnov omnibus test, we conclude now that the deviation between the medians (and thus the CDF) is statistically significant.

Note: The test can be generalized to the comparison of K ($K \geq 2$) medians. The test statistic follows a CHI-SQUARED (χ^2) distribution with a $(K-1)$ degrees of freedom. In our case, we obtain (for $K = 2$) $Z^2 = (2.29734)^2 = 5.27778 = \chi^2$. The « p-value » are exactly the same.

5.2 The Wilcoxon-Mann-Whitney test (U-test)

The equivalent nonparametric of t-test is the U-test. It is almost as powerful as the t-test is the Gaussian assumption is observed; it is more powerful as soon as we observe a deviation from this assumption. Compared with the median test above, it is also more powerful because it characterizes the differences between the values each other, and not only the differences of the values from the median. The U-test uses the rank sums of the two samples (<http://www.itl.nist.gov/div898/handbook/prc/section3/prc35.htm>).

We add the MANN-WHITNEY COMPARISON component (NONPARAMETRIC STATISTICS tab) into the diagram. Then, we click on the VIEW menu.

The screenshot shows the TANAGRA 1.4.27 interface with the 'Mann-Whitney Comparison 1' component selected. The 'Parameters' section shows 'Sort results: no'. The 'Results' table is as follows:

Value	Examples	Average	Rank sum	Rank mean	Mann-Whitney U	U
new	10	42.1400	69.5	6.9500	14.50000	14.50000
old	10	43.2300	140.5	14.0500	50.00000	50.00000
All	20	42.6850	210.0	10.5000	174.73684	174.73684

Additional results shown:

Z	2.68557
P(> Z)	0.00724

Computation time : 0 ms.
Created at 27/08/2008 06:13:40

The 'Components' panel at the bottom shows various statistical tests, with 'Mann-Whitney Comparison' highlighted in the 'Nonparametric statistics' tab.

The rank sum of the first group (MACHINE = NEW), which is the reference group, is $S_1 = 69.5$; for the second group, we obtain $S_2 = 140.5$. The test statistic is

$$U = S_1 - \frac{n_1(n_1 + 1)}{2} = 69.5 - \frac{10(10 + 1)}{2} = 14.5$$

TANAGRA provides $E(U) = 50.0$ and $V(U) = 174.73684$ under the null hypothesis. The standardized test statistic is computed as follows

$$|Z| = \frac{|U - E(U)|}{\sqrt{V(U)}} = \frac{|14.5 - 50.0|}{\sqrt{174.73684}} = 2.68557$$

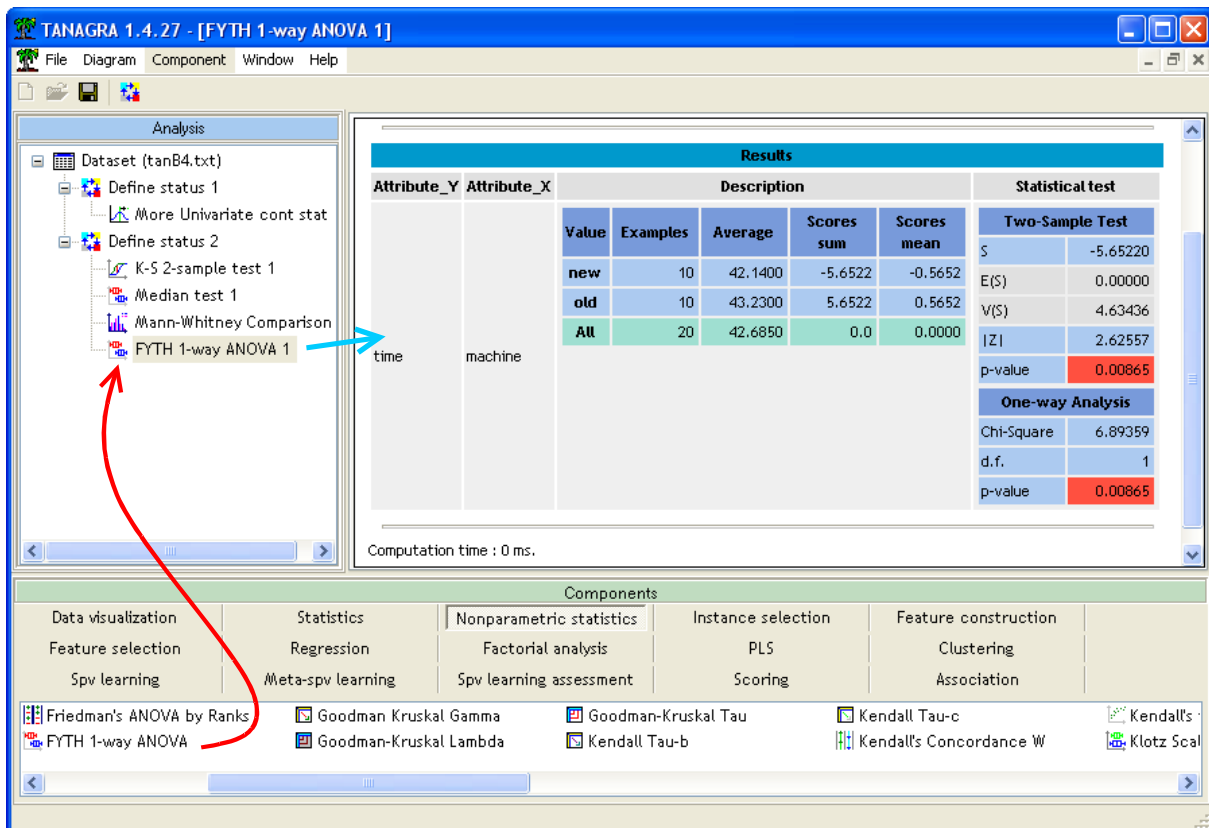
Then we obtain the p-value of the test: $p = 0.00724$.

The deviation between the CDF which seems obvious above (Figure 1) is clearly corroborated by the statistical test now. In our context, the U-test seems the most adapted according to the preliminary checking recommended on our reference website⁴.

5.3 Other nonparametric location tests

The Fisher-Yates-Terry-Hoeffding (FYTH) test and the Van der Waerden test are variants of the U-test. They are more adapted if the distribution of values is close to Gaussian distribution. They are also based on the rank of the data, but these last one are transformed into "normal scores" from which we compute the conditional sums (see http://en.wikipedia.org/wiki/Van_der_Waerden_test).

We add the FYTH 1-WAY ANOVA component (NONPARAMETRIC STATISTICS tab) into the diagram. We obtain the following results.



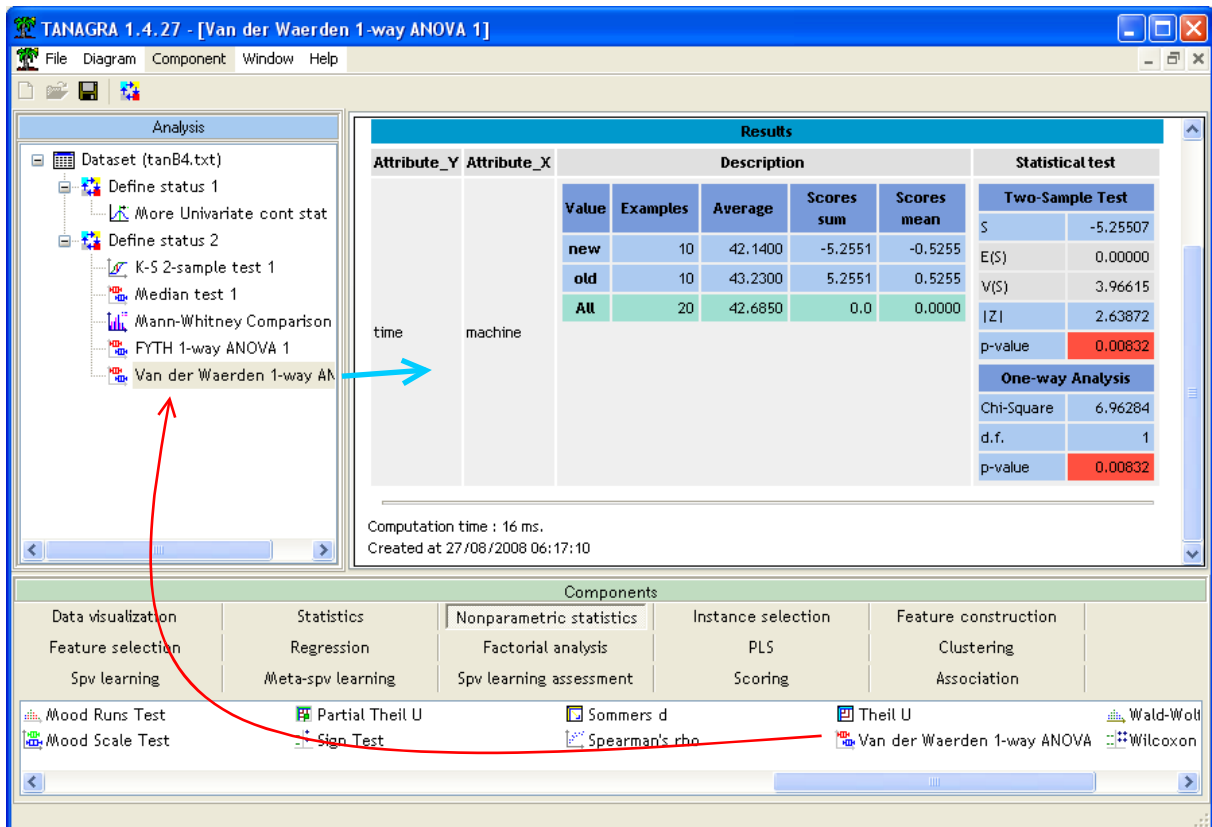
⁴ http://www.stat.psu.edu/online/development/stat500_spss/lesson10/lesson10_03.html

The test statistic, the scores sum of the reference group is $S = -5.6522$. Under the null hypothesis, we have $E(S) = 0$ and $V(S) = 4.63436$, we can compute the standardized test statistic

$$|Z| = \frac{|S - E(S)|}{\sqrt{V(S)}} = \frac{|- 5.6522|}{\sqrt{4.63436}} = 2.62557$$

The p-value of the test is $p = 0.00865$. The results are similar to those of the U-test.

The Van der Waerden test is slightly different to the FYTH test according to the computation of the "scores". We add the VAN DER WAERDEN 1-WAY ANOVA component (NONPARAMETRIC STATISTICS tab) into the component. The results are very similar to those of the FYTH test.



6 Conclusion

The MEDIAN TEST, FYTH 1-WAY ANOVA and VAN DER WAERDEN tests are operational even if the number of groups is greater than 2 ($K > 2$). In this case, the Z statistic is hidden. Only the analysis of variance on "scores" (χ^2 statistic) is shown.

This is not possible for the U-test. It is then more adapted to use the Kruskal-Wallis test (KRUSKAL-WALLIS 1-WAY ANOVA component) which is the variant of the U-test for ($K > 2$) groups.