

1 Subject

Tests for comparing two related samples.

Dependent samples, also called related samples or correlated samples, occur when the response of the n^{th} person in the second sample is partly a function of the response of the n^{th} person in the first sample. There are several common forms of sample dependency¹. (1) Before-after and other studies in which the same people are surveyed at different points in time, including panel studies. (2) Matched-pairs studies in which each of the subjects of the study is paired with each of those in a comparison group on the basis matching factors (e.g. age, sex, income, etc.). (3) The pairs can simply be inherent in the situation we are trying to analyze. For instance, one tries to compare the time spent watching television by the man and woman within a couple. The blocks are naturally households. Men and women should not be considered as independent observations.

The aim of tests for related samples is to exclude from the analysis the within-group variation. The calculation of the differences is realized within each pair of subjects. In this tutorial, we show how to implement 3 tests for two related samples. Two of them are non-parametric (sign test and Wilcoxon matched-pairs ranks test), the last one is the parametric t-test for related samples.

2 Dataset

The dataset come from the Pr Richard Lowry course website². The approach, formulas and calculations related to this dataset are detailed on the website. The data file is available on line³.

Suppose that 16 students in an introductory statistics course are presented with a number of questions concerning basic probabilities. In each instance, the question takes the form "What is the probability of such-and-such?". However, the students are not allowed to perform calculations. Their answers must be immediate, based only on their raw intuitions. The instructor of the course is particularly interested in student's responses to two of the questions, which we will designate as question A and question B. He reasons that if students have developed a good, solid understanding of the basic concepts, they will tend to give higher probability ratings for question A than for question B; whereas, if they were sleeping through that portion of the course, their answers will be mere shots in the dark and there will be no overall tendency one way or the other.

3 The non-parametric tests for two related samples

3.1 Importing the dataset

The easiest way to handle the dataset is to open the file into Excel spreadsheet. We select the data range and we click on the TANAGRA / EXECUTE TANAGRA menu which is installed with the TANAGRA.XLA add-in⁴.

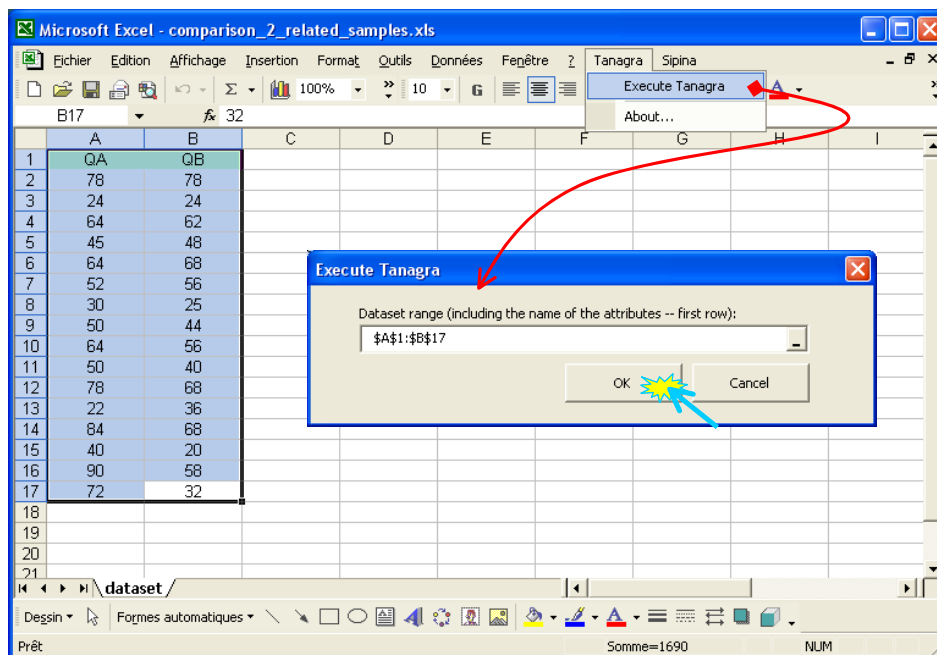
¹ <http://faculty.chass.ncsu.edu/garson/PA765/mcnemar.htm>

² <http://faculty.vassar.edu/lowry/ch12a.html>; the analyzed problem is largely explained on the site.

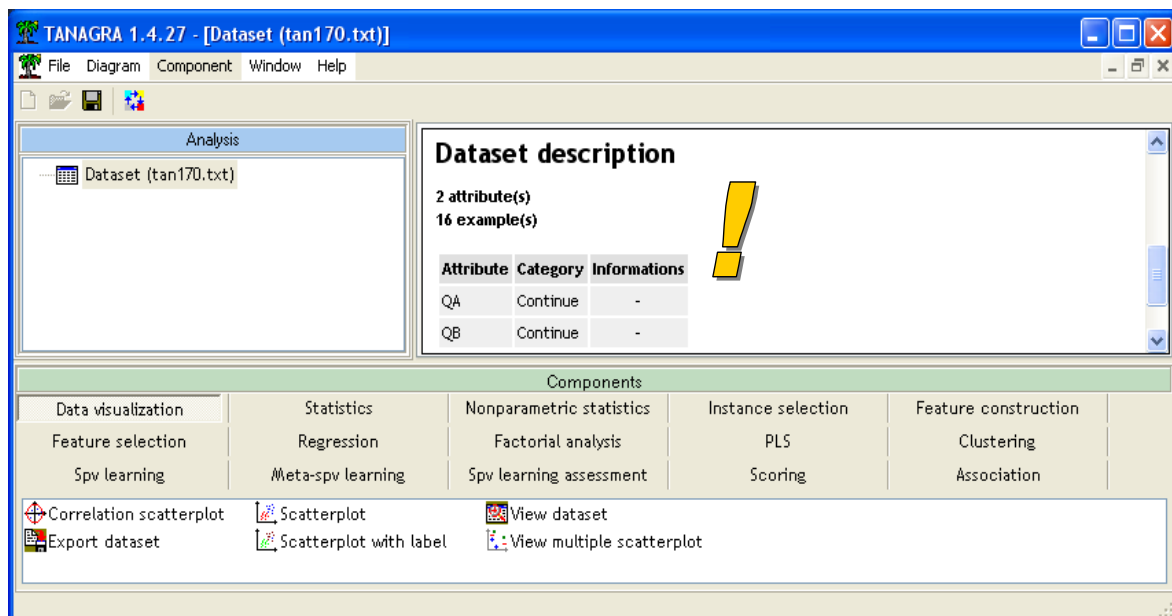
³ http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/comparison_2_related_samples.xls

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

A dialog box appears. We check the selection. Then we validate by clicking on the OK button.



TANAGRA is automatically launched. A new diagram is created. We have 16 instances and 2 variables (QA and QB).

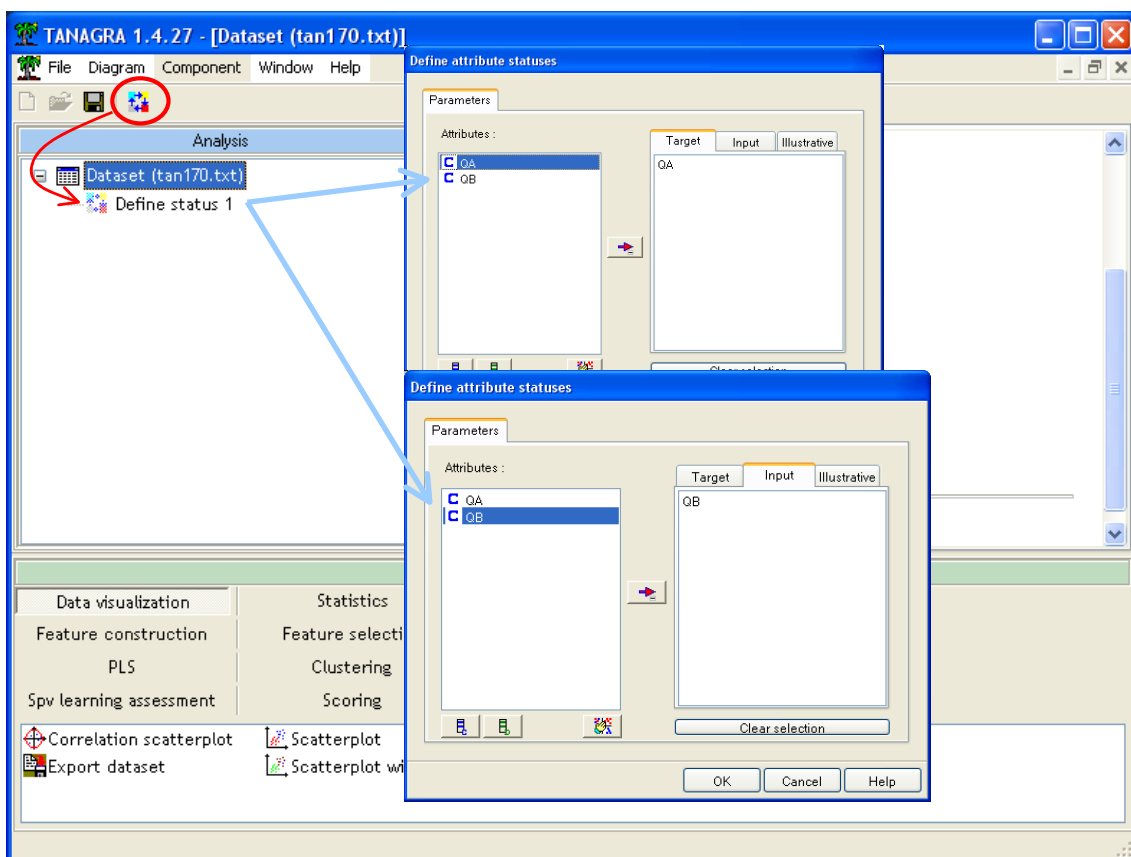


All tests that we present in this tutorial are based on a synthetic variable, constructed from the difference between the values of QA and QB for each observation of the dataset i.e. $D = QA - QB$. They differ in how to exploit this new variable D.

3.2 Sign test

The sign test uses only the sign of D. Under the null hypothesis, the theoretical probability of obtain a positive value for D is 0.5^5 . We try to find if we obtain an observed value of D which deviates significantly to 0.5.

We insert the DEFINE STATUS component from the shortcut into the toolbar. We set QA as TARGET, QB as INPUT.



Then we add the SIGN TEST component (NONPARAMETRIC STATISTICS tab). We click on the VIEW menu to obtain the results.

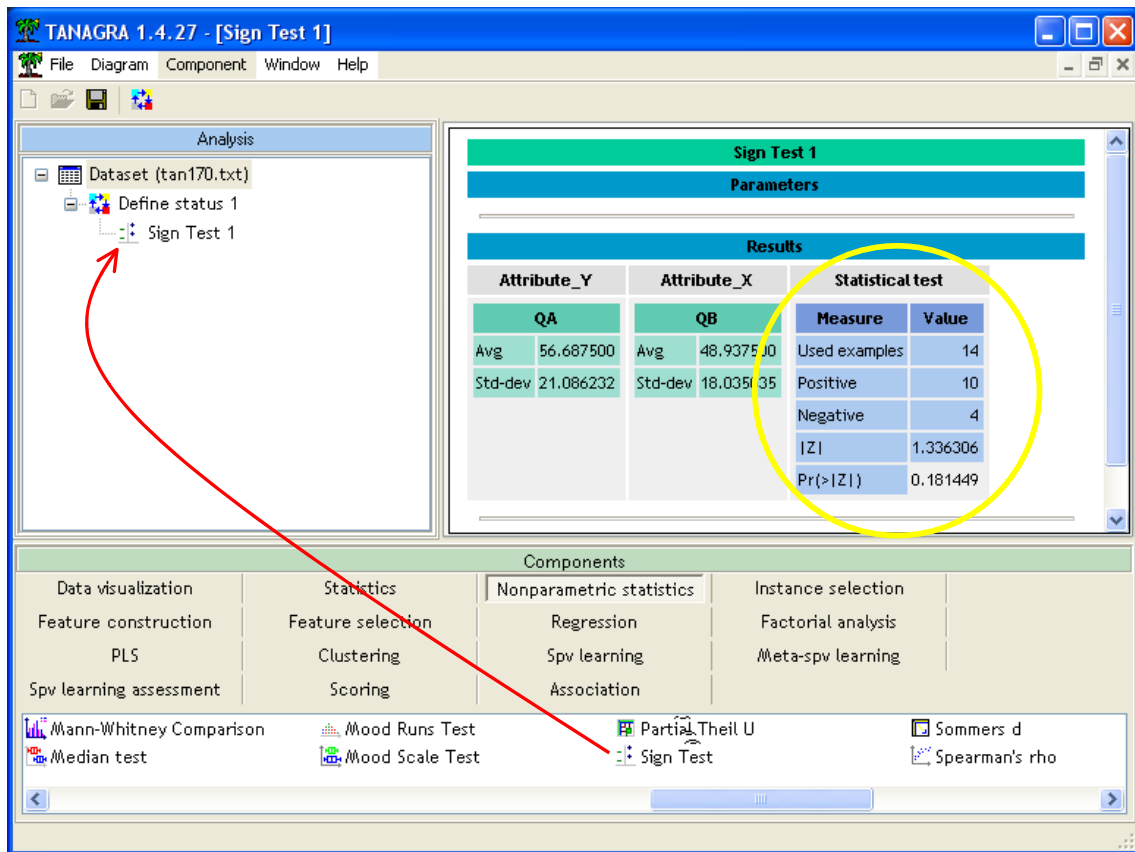
Initially, we have 16 instances. Only $n = 14$ of them are usable. They correspond to the instances for which $(D \neq 0)$. There are $S = 10$ positives values ($D > 0$) and $n - S = 4$ negative ones ($D < 0$).

The standardized test statistic, with the continuity correction, is

$$|Z| = \frac{|2S - n| - 1.0}{\sqrt{n}} = \frac{|2 \times 10 - 14| - 1.0}{\sqrt{14}} = 1.336306$$

The p-value of the test is $p = 0.181449$. At the 5% significance level, the difference between QA and QB is not significant.

⁵ <http://udel.edu/~mcdonald/statsign.html> ; http://en.wikipedia.org/wiki/Sign_test



But this test is not really sensitive. It is rather conservative i.e. it often favors the null hypothesis (the deviation is not significant). Then we must take with caution this result, we must use more powerful tests.

3.3 Wilcoxon signed-rank test

This test uses the relative importance of the deviations for the construction of the test statistic. It is more sensitive to the importance of the deviations⁶.

We insert the WILCOXON SIGNED RANKS TEST component (NONPARAMETRIC STATISTICS tab) into the diagram.

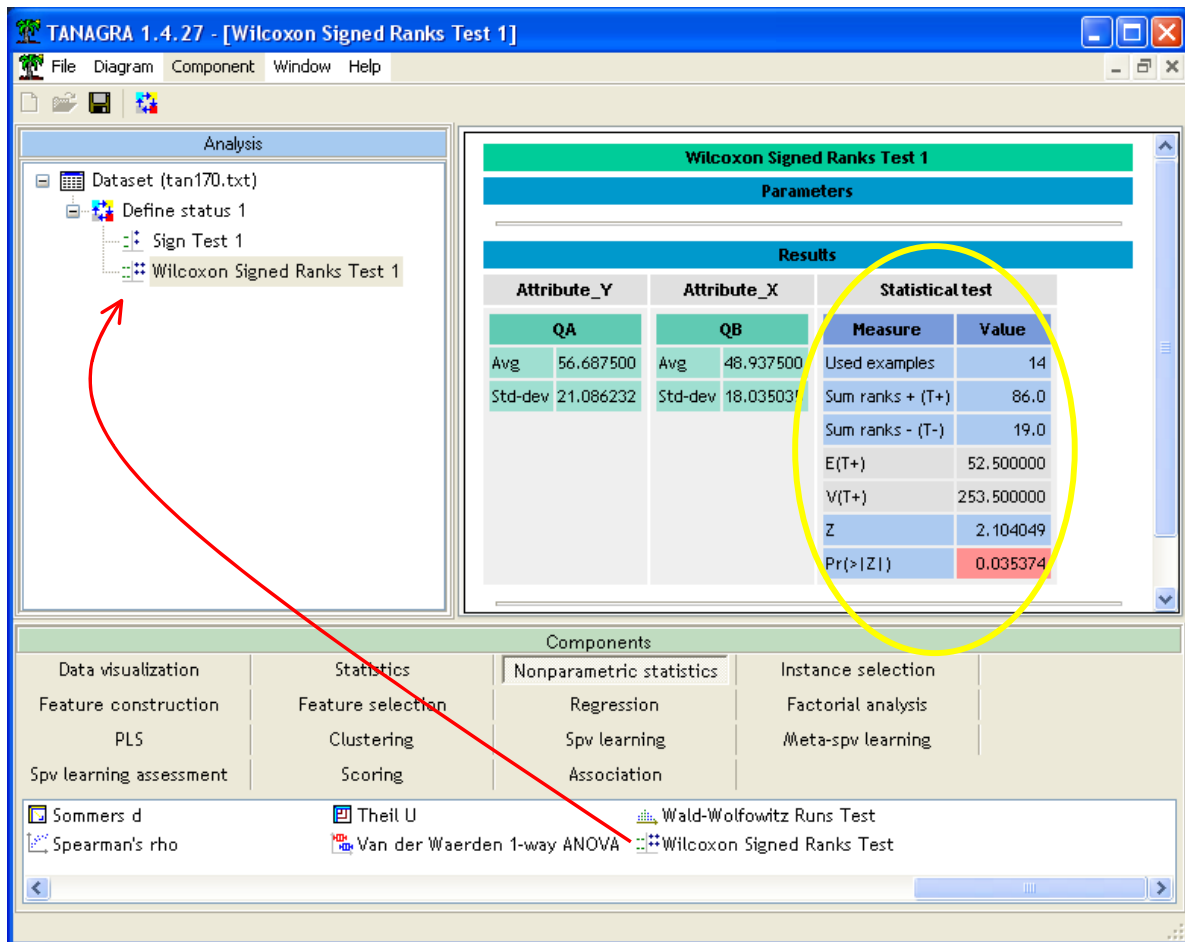
Here also the number of usable instances is $n = 14$ ($D \neq 0$). The sum of the ranks of the positive deviations is $T^+ = 86$. Its mean and variance under the null hypothesis are $E(T^+) = 52.5$ and $V(T^+) = 253.5$. The standardized test statistic for a bilateral test, without the continuity correction, is⁷

$$|Z| = \frac{|T^+ - E(T^+)|}{\sqrt{V(T^+)}} = \frac{|86 - 52.5|}{\sqrt{253.5}} = 2.104049$$

⁶ <http://udel.edu/~mcdonald/statsignedrank.html>

⁷ See <http://faculty.vassar.edu/lowry/ch12a.html> for the details of the calculations.

The p-value of the test is $p = 0.035374$. Now, because we make a better use of the available information, we observe a significant deviation between QA and QB.



4 T-test for related samples

The t-test for related samples uses the values of the differences ($D = QA - QB$). Under the null hypothesis, the mean of D is equal to 0. We come down to a one sample t-test scheme⁸.

4.1 Normality test for D

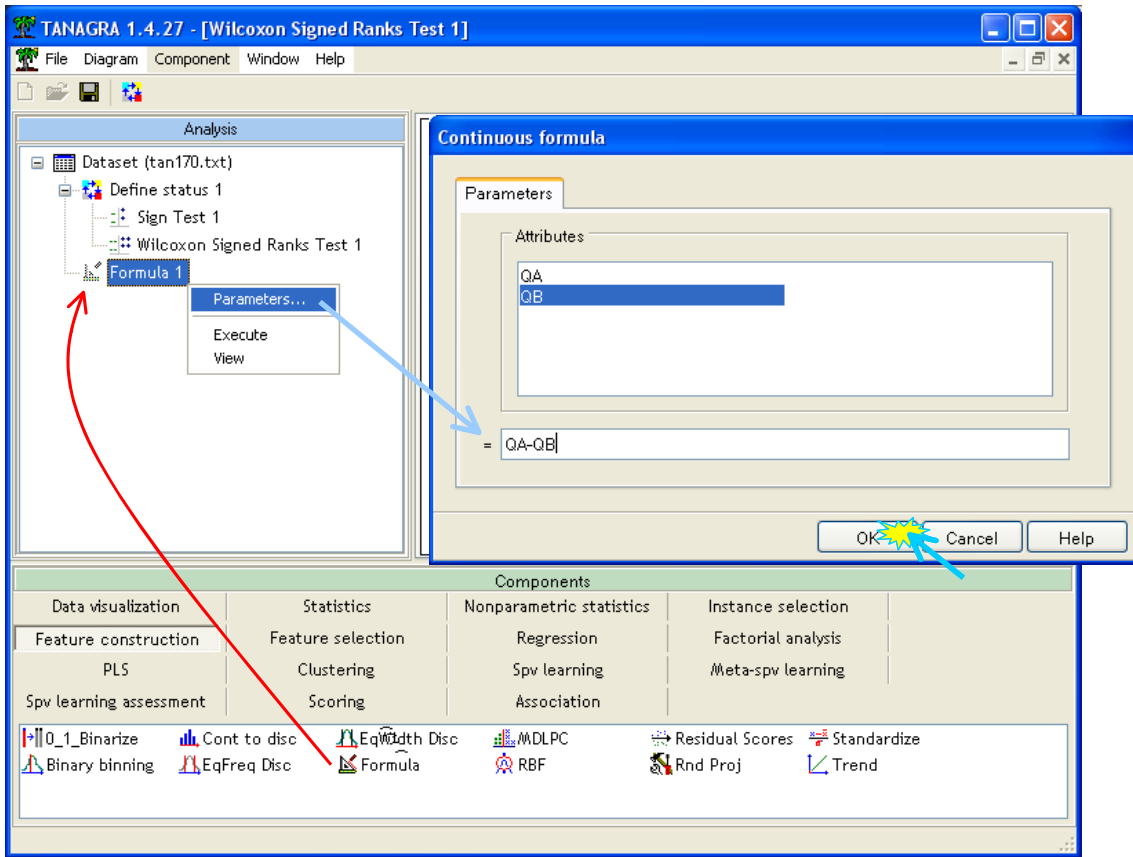
For the test to be valid, the distribution of D must be compatible with the Gaussian distribution.

We will create explicitly the synthetic variable D. We want to control its distribution using the usual normality tests⁹.

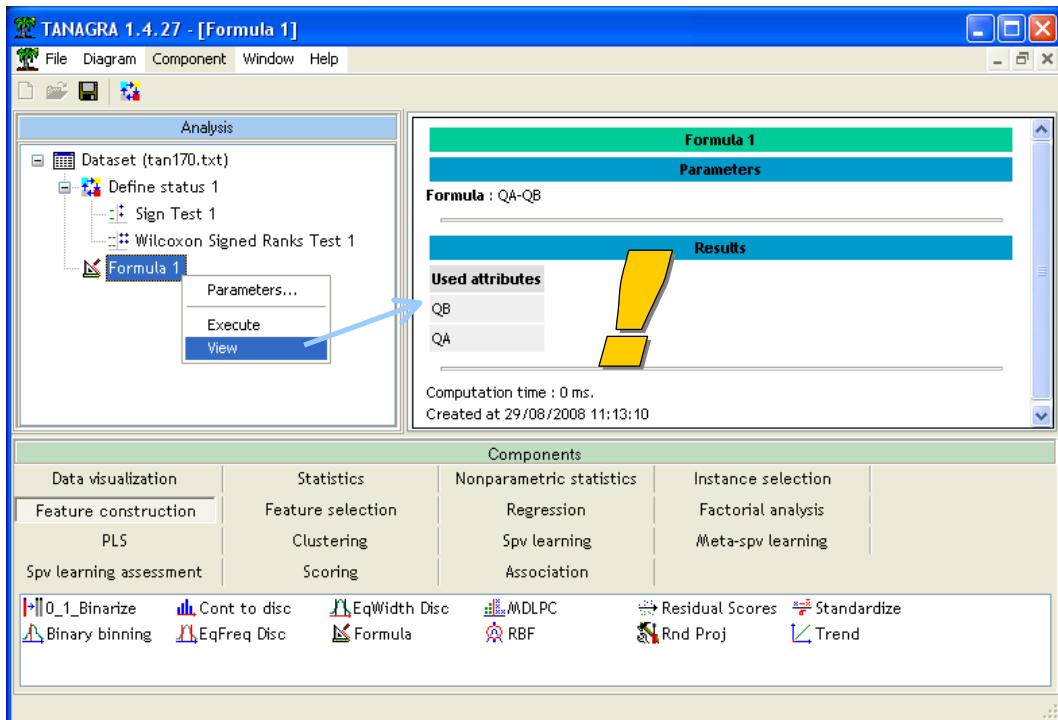
We insert the FORMULA component into the diagram (FEATURE CONSTRUCTION tab). We click on the PARAMETERS contextual menu. We set the formula for D.

⁸ <http://www.itl.nist.gov/div898/handbook/eda/section3/eda352.htm>

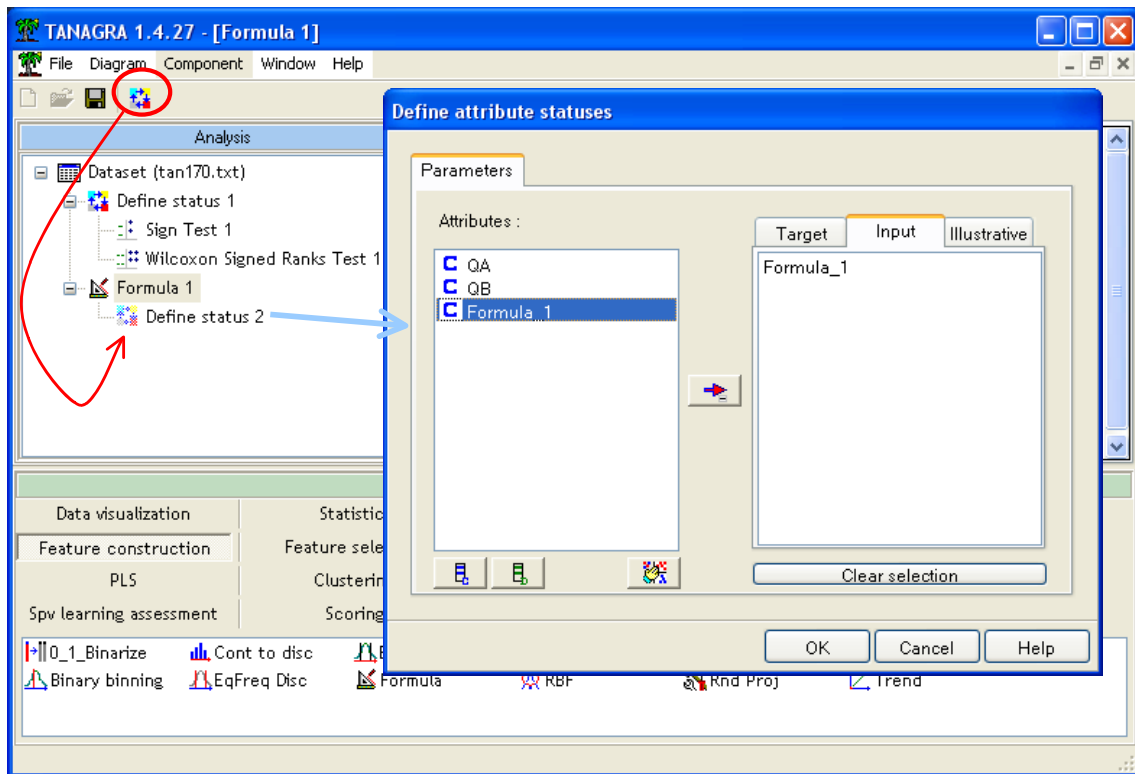
⁹ http://en.wikipedia.org/wiki/Normality_test



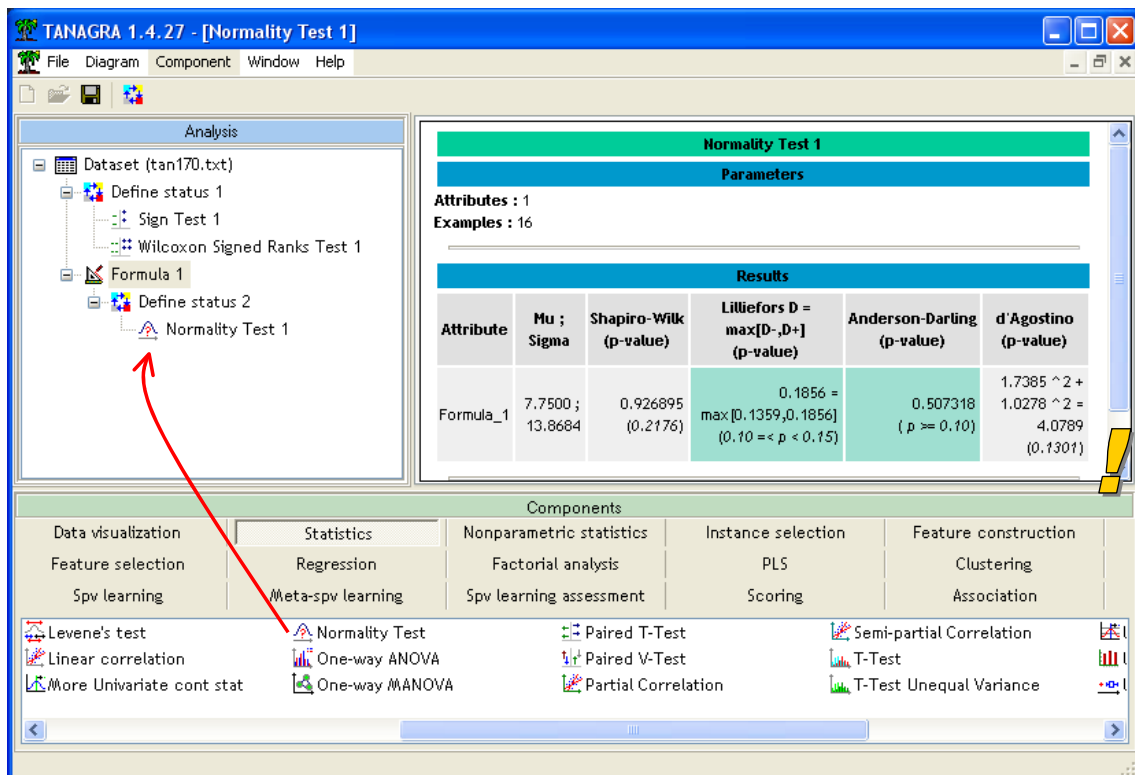
We click on the OK button in order to validate the formula. Then we click on the VIEW menu. Tanagra creates a new column named "FORMULA_1". We can use it in the subsequent part of the diagram.



We insert a new DEFINE STATUTS component. We set the new column as INPUT.



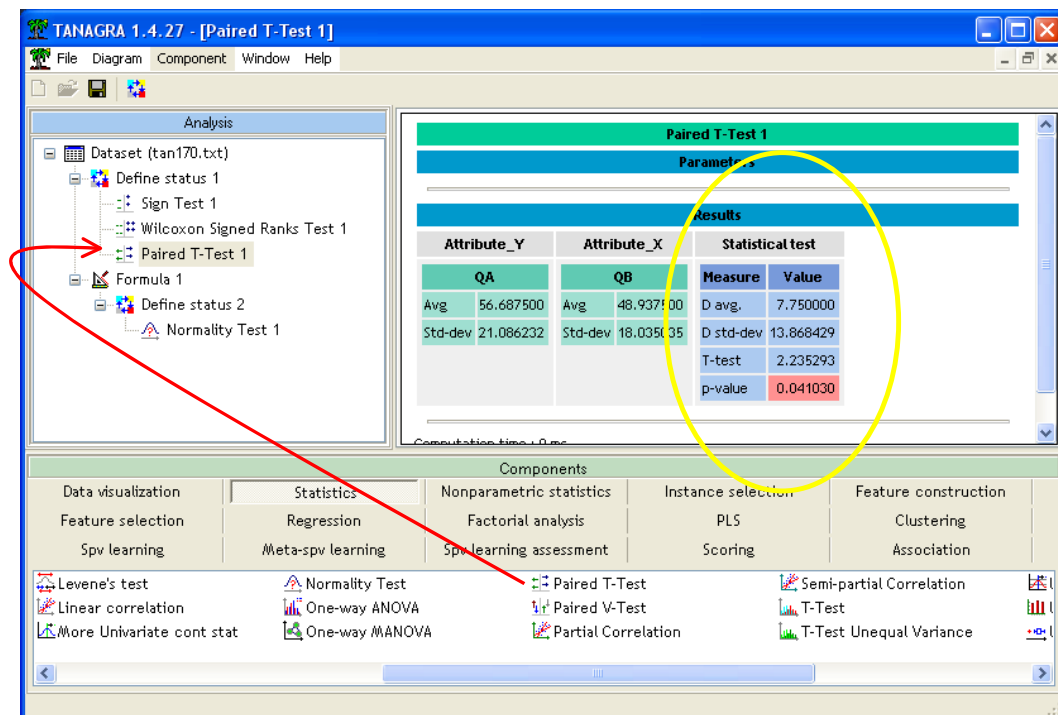
Then we insert the NORMALITY TEST component (STATISTICS tab) into the diagram.



It seems that the distribution of D is compatible with the Gaussian distribution. On all used tests, at the 5% significance level, we cannot reject the normality hypothesis.

4.2 Paired samples t-test

We can implement the t-test now. We add the PAIRED T-TEST component (STATISTICS tab) below the DEFINE STATUS 1. We click on the VIEW menu.



The mean of D is $\bar{D} = 7.75$, its standard deviation $s_D = 13.868429$. We can compute the test statistic t . We use the null deviations here i.e. $n = 16$.

$$t = \frac{\bar{D}}{\frac{s_D}{\sqrt{n}}} = \frac{7.75}{\frac{13.868429}{\sqrt{16}}} = 2.235293$$

Under the null hypothesis H_0 , the distribution of t is a Student distribution with $(n-1) = 16 - 1 = 15$ degrees of freedom. The p -value of the test is $p = 0.04103$. This result is close to the one of Wilcoxon signed rank test.

5 Graphical representation

It seems that the deviation between QA and QB is significant according to the Wilcoxon signed rank test and the Paired samples t-test. But, a simple graphical representation can give us the same conclusion without complicated calculations. It is more preferable in many situations because it provides more information about the dataset (e.g. outliers).

About our context, we can use, among other, a scatter plot¹⁰. Under the null hypothesis, we should see the collection of points more or less aligned on the main diagonal.

¹⁰ http://en.wikipedia.org/wiki/Scatter_plot

On our dataset, we note that the points are in majority under the main diagonal (Figure 1). The values are different and, in light of this graphical representation, one can even say that QA tends to be larger than QB.

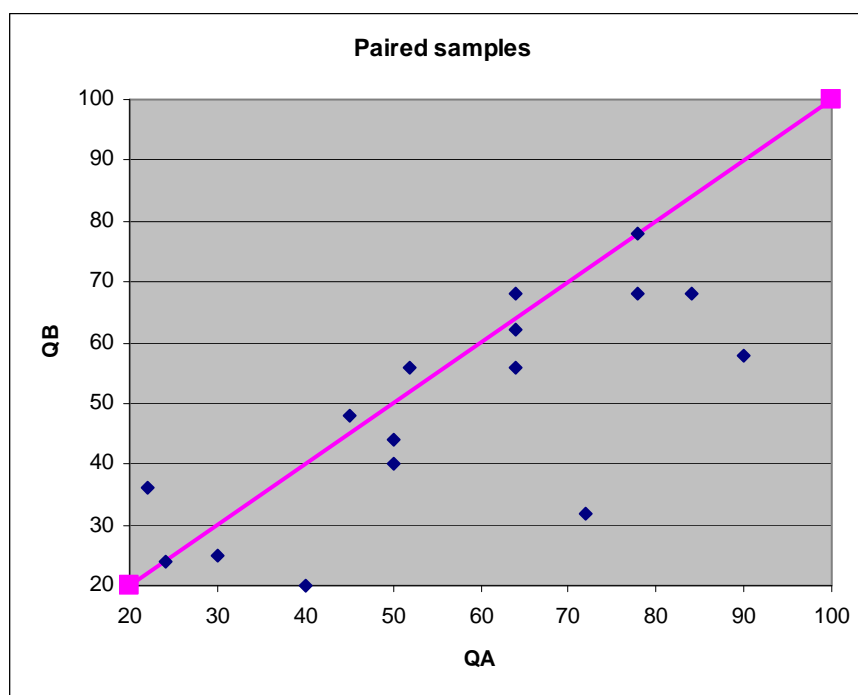


Figure 1 – Scatter plot QA vs. QB