

## 1 Subject

**Parametric hypothesis testing for comparison of two or more populations. Independent and dependent samples.**

The tests for comparison of population try to determine if  $K$  ( $K \geq 2$ ) samples come from the same underlying population according to a variable of interest ( $X$ ). We talk parametric test when we assume that the data come from a type of probability distribution. Thus, the inference relies on the parameters of the distribution. For instance, if we assume that the distribution of the data is Gaussian, the hypothesis testing relies on mean or on variance.

We handle univariate test in this tutorial i.e. we have only one variable of interest. When we want to analyze simultaneously several variables, we talk about multivariate test.

Samples are independent if the response of the  $n^{\text{th}}$  person in the second sample is not a function of the response of the  $n^{\text{th}}$  person in the first sample. Independent samples are also called uncorrelated or unrelated samples. Dependent samples, also called related samples or correlated samples, occur when the response of the  $n^{\text{th}}$  person in the second sample is partly a function of the response of the  $n^{\text{th}}$  person in the first sample (e.g. before-after scheme; matched-pairs studies; etc.)<sup>1</sup>.

The test that we describe in this tutorial can be used to compare process (e.g. is it that two machines produce the same diameter bolts), but it enables also to test the connection that can exist between a categorical variable and a quantitative variable (e.g. is that women drive slower on average than men on the roads).

## 2 Dataset

The CREDIT\_APPROVAL.XLS<sup>2</sup> data file describes 50 households consisting of married couples, both actives, which asked for a credit to a bank. The available variables are:

Variable	Description
Sal.Homme	Logarithm of the male salary
Sal.Femme	Logarithm of the female salary
Rev.Tete	Logarithm of the income per head in the household
Age	Logarithm of the male age
Acceptation	Acceptation of the credit
Garantie.Supp	Additional guarantee is required
Emploi	Type of job held by the reference person of the household

All the continuous variables are potentially a variable of interest; the categorical variables are used to define the sub-populations.

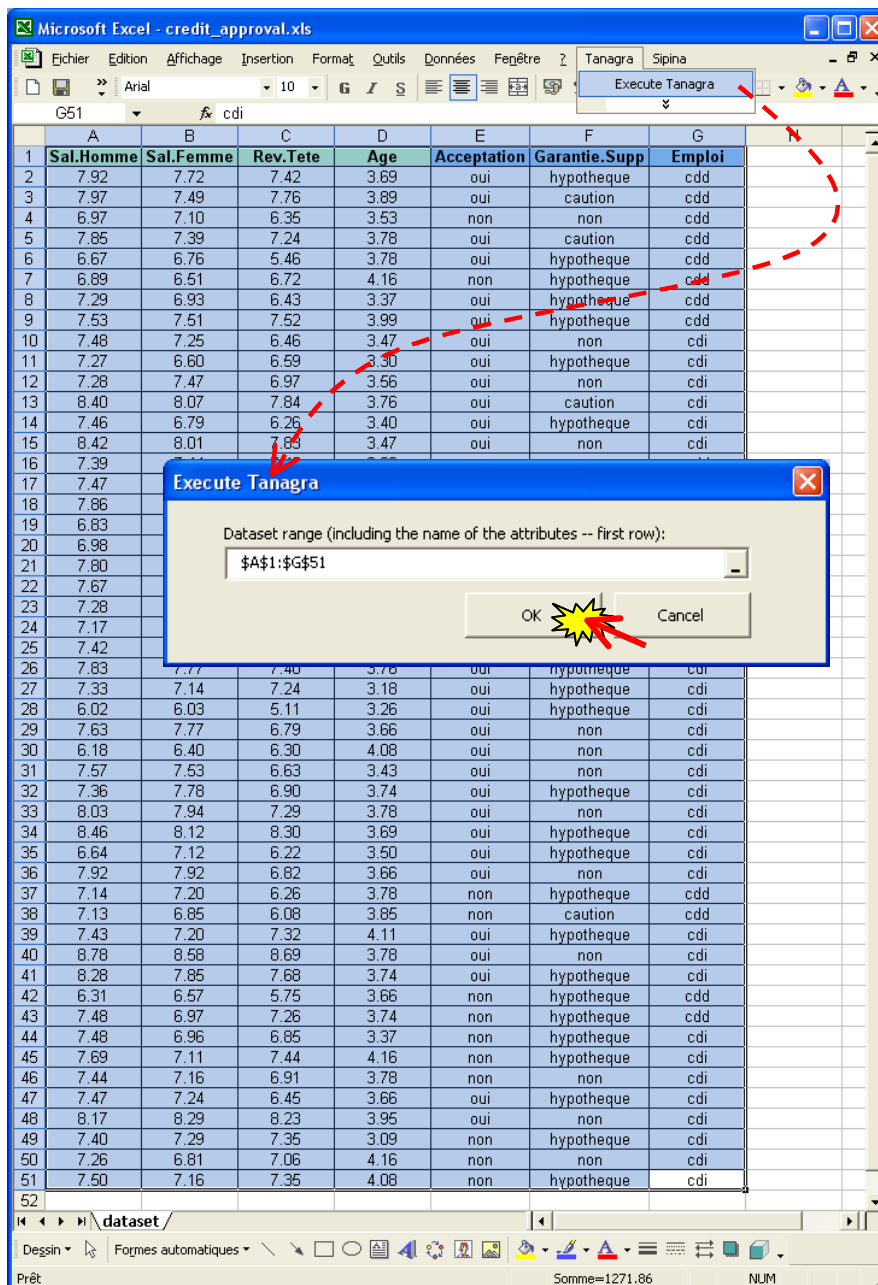
<sup>1</sup> See <http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm>

<sup>2</sup> [http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/credit\\_approval.xls](http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/credit_approval.xls)

### 3 Descriptive statistics and normality test

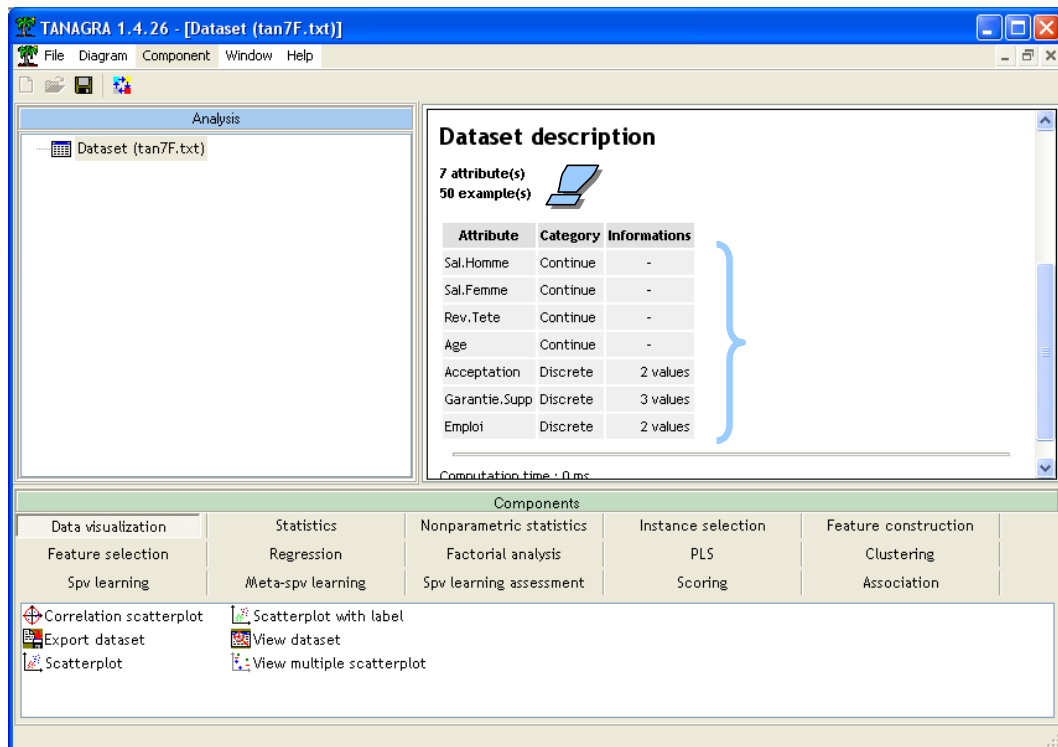
#### 3.1 Importing the data file into Tanagra

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<sup>3</sup>. After we checked the coordinates of the selected cells, we click on OK button.



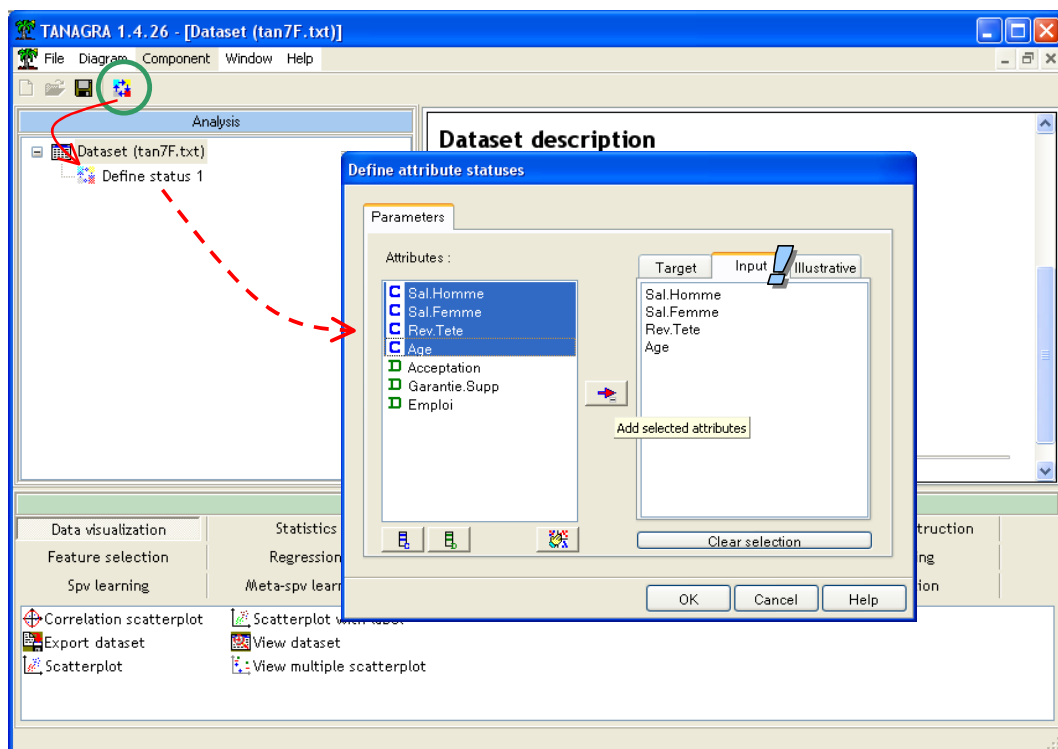
Tanagra is automatically launched. A new diagram is created. We verify that the data set has **50 observations** and **7 variables**.

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

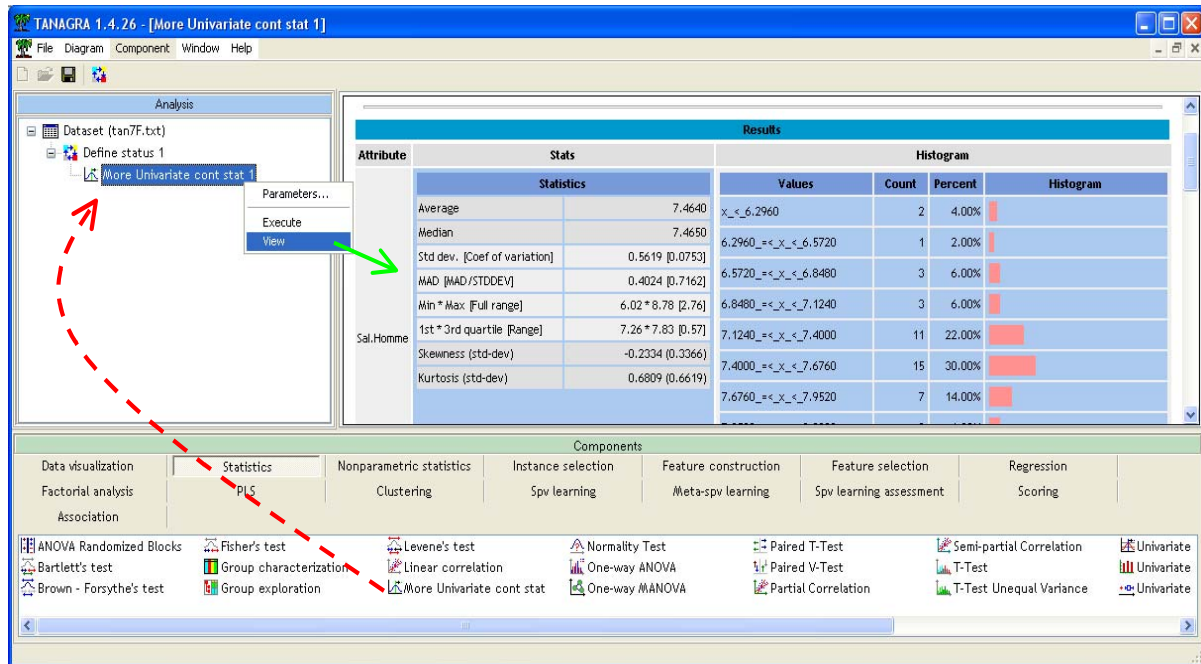


### 3.2 Descriptive statistics

In a first step, we want to describe our dataset by computing some descriptive statistics indicators on quantitative variables. We add the DEFINE STATUS component into the diagram. We set Sal.Homme, Sal.Femme, Rev.Tete and Age as INPUT.



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



With such tools, we try primarily to detect anomalies (e.g. skewed or bimodal distribution, the existence of outliers, etc.). It seems that there is no need to worry too much here.

### 3-3 Normality test

The tests we present in this tutorial assume Gaussian distribution of variables of interest. Certainly they are more or less robust to this assumption. However, to ensure the viability of the results (and show you how to do), we decided to conduct this checking<sup>4,5</sup>.

A first empirical evaluation was possible with the previous component. We know that the ratio between the MAD (mean absolute deviation) and the SD (standard deviation) is approximately 0.8 if the sample comes a Gaussian population<sup>6</sup>. Thus if the ratio differs from this value, we suspect that the underlying distribution is not Gaussian (the inverse is not true, if the ratio is 0.8, it does not mean that the underlying distribution is necessarily Gaussian).

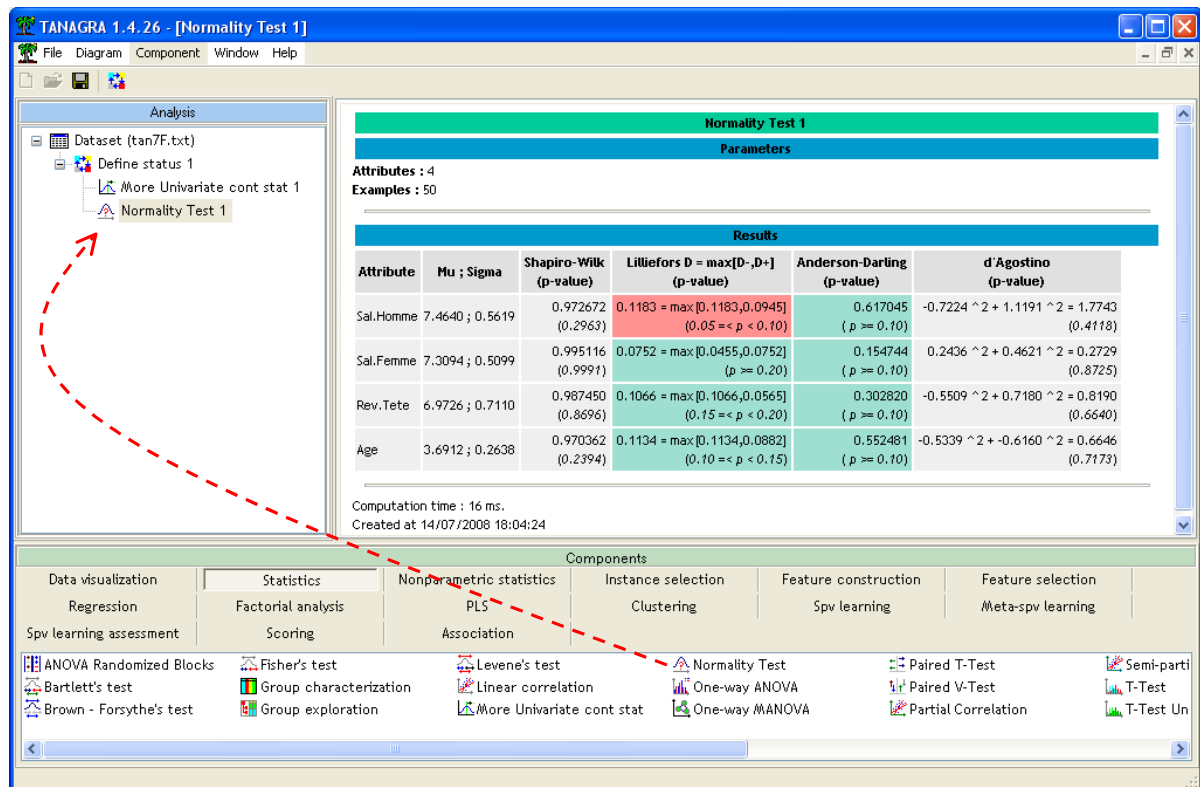
We mainly use this empirical rule for detecting the strong deviations. On our dataset, it seems that none of the variables diverge strongly to the Gaussian distribution according this rule: we 0.7162 for « Sal.Homme »; 0.7846 for « Sal.Femme », 0.7986 for « Rev.Tete » and 0.7895 pour « Age ». Apart from "Sal.Homme, we are even surprisingly close to the reference value.

<sup>4</sup> See <http://data-mining-tutorials.blogspot.com/2008/11/normality-test.html> for more details about the normality test.

<sup>5</sup> In the context of comparison of population, performing a normality test of the whole sample before the comparison is not always relevant. Indeed, if the alternate hypothesis is true, the conditional distribution (in the sub-population) is actually Gaussian, but the distribution of the whole sample is not Gaussian (e.g. a bimodal distribution can be a characteristic a significant difference between the means of the sub populations). It is more suitable to perform the normality test within each subgroup.

<sup>6</sup> More precisely  $\sqrt{\frac{2}{\pi}} \approx 0.798$ . See [http://en.wikipedia.org/wiki/Mean\\_absolute\\_deviation](http://en.wikipedia.org/wiki/Mean_absolute_deviation)

Now we proceed to a more formal evaluation using normality tests. We add the NORMALITY TEST (STATISTICS tab) into the diagram. We click on the VIEW menu.



At the 5% significance level, all the variables seem compatible to the Gaussian distribution.

## 4 Tests for independent samples

### 4.1 Comparison of two means with equal variances

We want to compare the female salary according to the credit acceptance. In other words, we want to know if the credit acceptance relies on the salary of the female<sup>7</sup>.

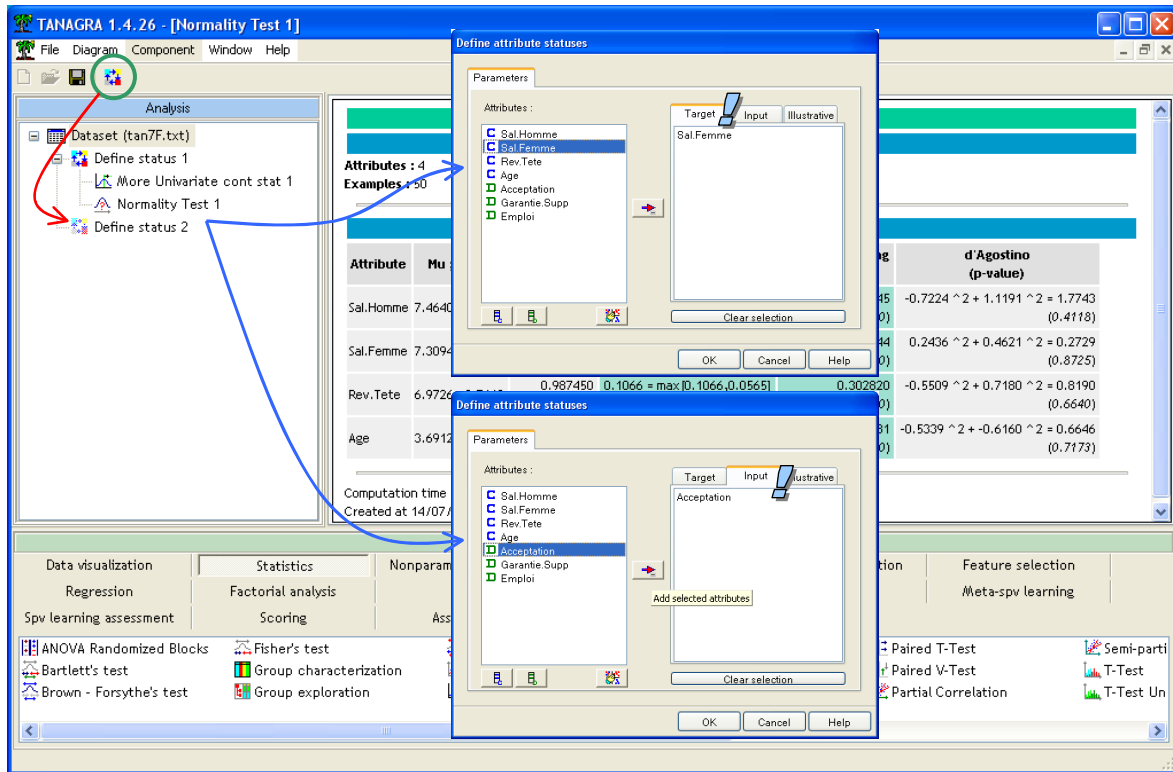
We are in a comparison of means scheme. We use the t-test of difference in means. We assume that the variances are the same in the two populations (acceptation of credit is yes or no)<sup>8,9</sup>.

We add the DEFINE STATUS component into the diagram using the shortcut into the toolbar. We set SAL.FEMME as TARGET, ACCEPTATION as INPUT.

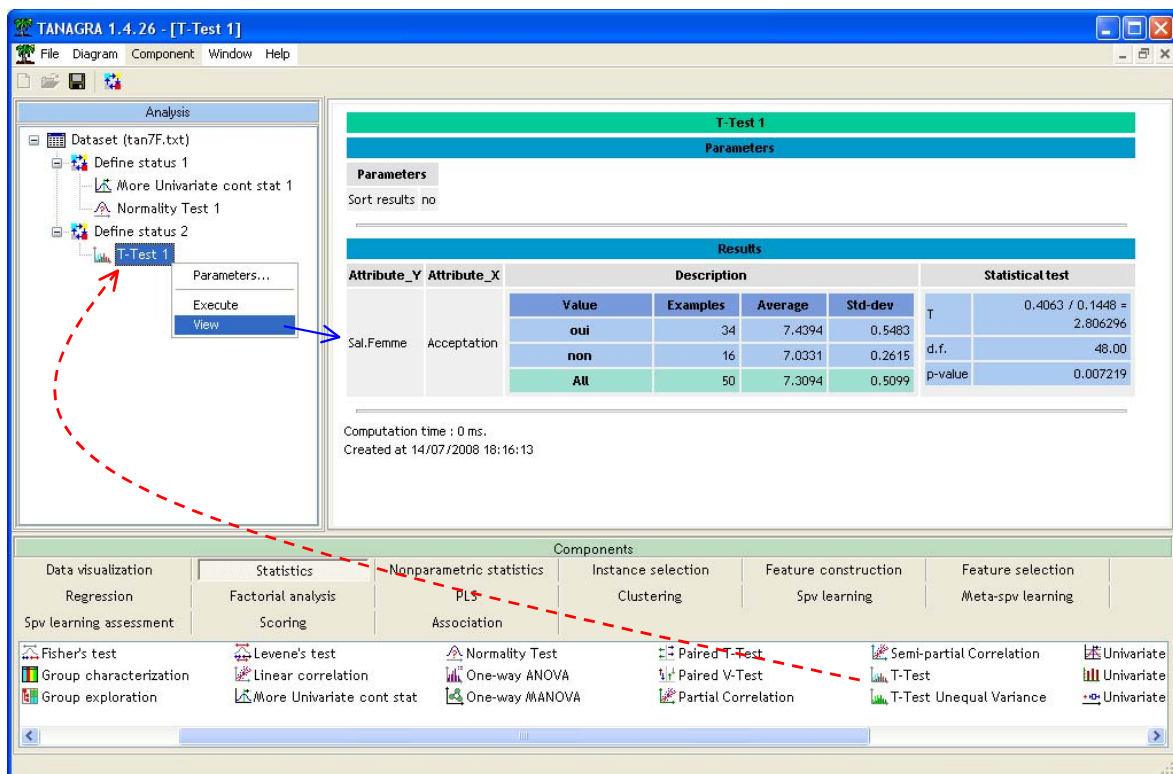
<sup>7</sup> The opposite interpretation is not relevant. The credit acceptance cannot influence the salary of the female.

<sup>8</sup> <http://www.itl.nist.gov/div898/handbook/prc/section3/prc31.htm>

<sup>9</sup> [http://en.wikipedia.org/wiki/Student's\\_t-test](http://en.wikipedia.org/wiki/Student's_t-test)



Then, we insert the T-TEST component (STATISTICS tab). We click on the VIEW menu in order to obtain the results.

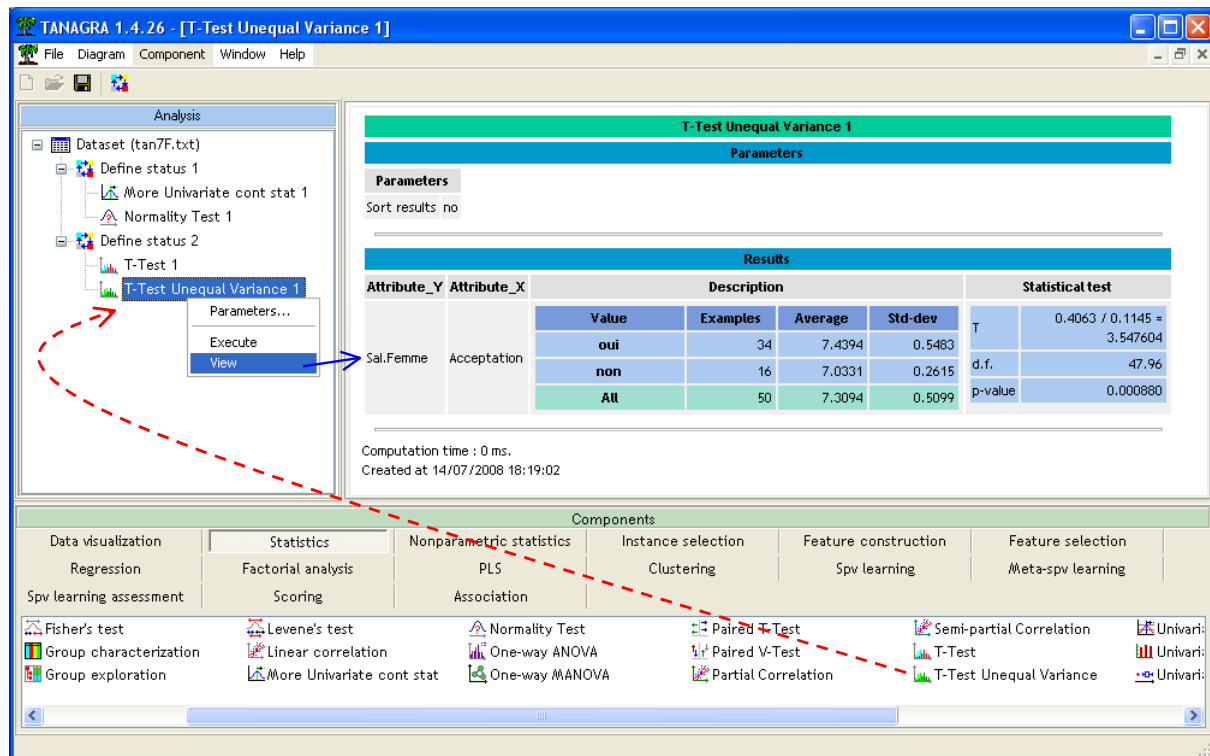


The t-statistic is  $t = 0.4063 / 0.1448 = 2.8063$ . The difference is significant at the 5% significance level since the p-value is 0.007219. The credit acceptance is related to the female salary in the household.

It must however take with caution this result. The conditional variances are quite different; they range from simple to double. And the numbers of instances are unbalanced. The condition of applicability of the test-statistic for equal variances does not seem satisfied. We must refine the analysis by introducing the t-test for unequal variances.

#### 4.2 Comparison of two means – Unequal variances

This version does not use the pooled variance for the estimation of the variance of the deviation. The degree of freedom is also corrected. We add the T-TEST UNEQUAL VARIANCE component (STATISTICS tab) into the diagram. We obtain the following results.

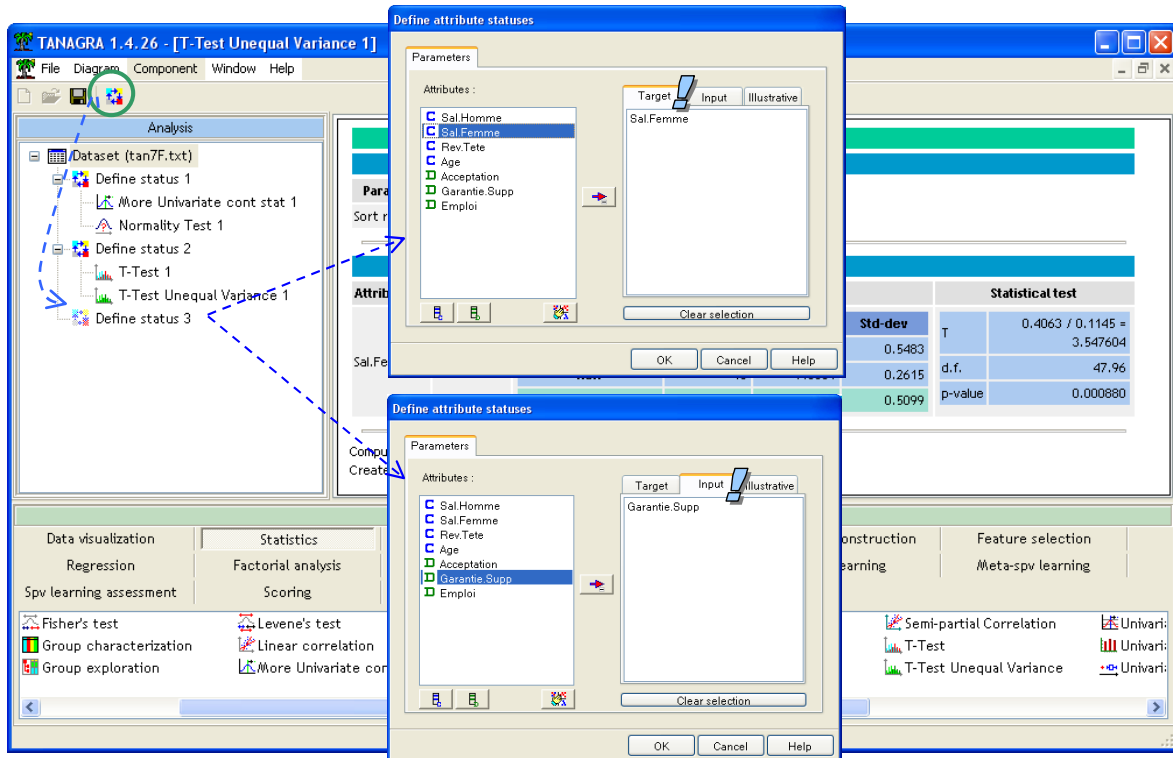


In comparison to the previous statistic, the numerator is the same, but the denominator is different. We thus obtain  $t = 0.4063 / 0.1145 = 3.5476$ . The degree of freedom is  $d.f. = 47.96$ , so if we round it to the nearest integer,  $d.f. \approx 48$ . The p-value of the test is lower  $p\text{-value} = 0.00088$ . The salary gap according to the acceptance of credit is confirmed.

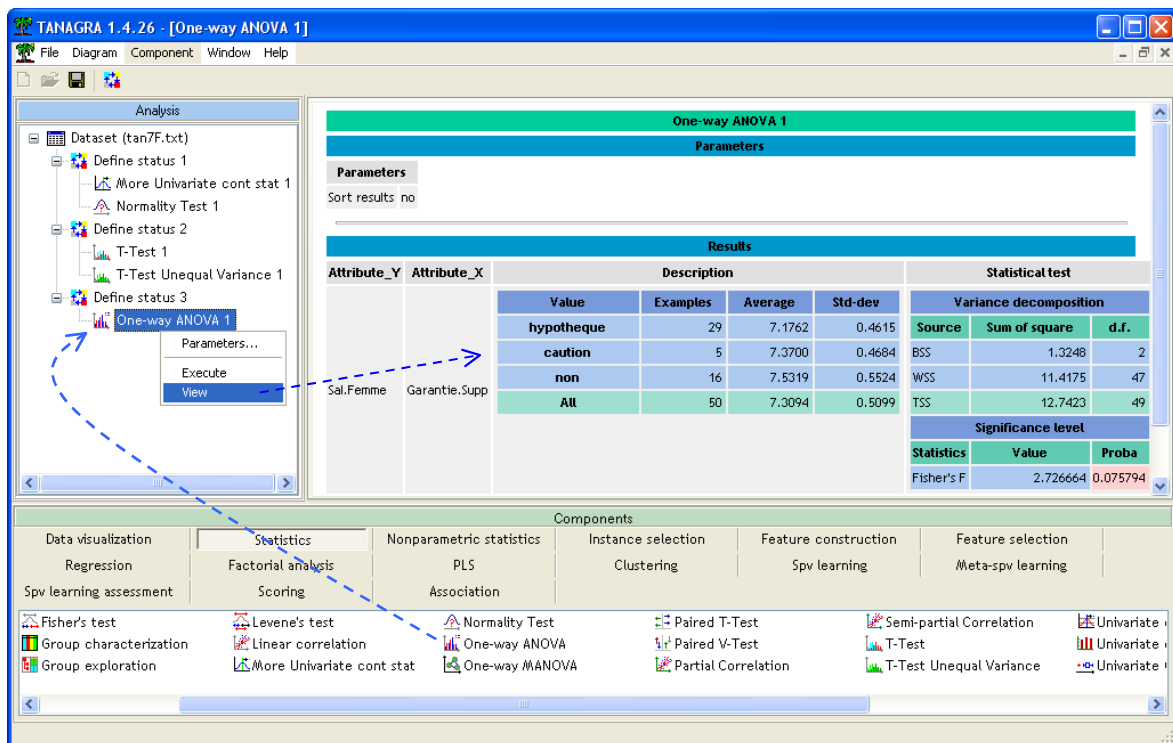
#### 4.3 Comparison of means for K populations – Equal variances (ANOVA)

We can see the one-way analysis of variance (ANOVA) as a generalization of the comparison of means for more than two populations. We want to compare women's wages in 3 groups defined by the type of additional warranty used by borrowers (GARANTIE.SUPP). Initially, we consider that the variances are identical in groups.

We add the DEFINE STATUS component into the diagram. We set SAL.FEMME as TARGET, GARANTIE.SUP as INPUT.



We insert the ONE-WAY ANOVA component (STATISTICS tab). We obtain the following results.



We obtain the conditional means and the conditional standard deviation. The F-statistic is  $F = 2.7267$ . The deviation between means seems not significant at the 5% level since the p-value of the test is  $p\text{-value} = 0.07579$ .



Even if the conditional standard deviation seems similar, the size of the groups is highly unequal in our dataset. We know that the test is very sensitive to the Behrens-Fisher problem. It is more careful to implement the test which does not rely on the pooled variance. We use the Welch's version.

#### 4.4 Comparison of means for K populations – Unequal variances

This version is adapted to the case where the variances are unequal within subgroups. We insert the WELCH ANOVA component (STATISTICS tab) into the diagram.

The screenshot shows the TANAGRA 1.4.26 interface. The 'Analysis' tree on the left includes 'Define status 3' with 'One-way ANOVA 1' and 'Welch ANOVA 1'. The 'Parameters' tab for 'Welch ANOVA 1' is active, showing 'Sort results: no'. The 'Results' tab displays a table with the following data:

Attribute_Y	Attribute_X	Description				Statistical test	
		Value	Examples	Average	Std-dev	Fisher's F	2,344622
Sal.Femme	Garantie.Supp	hypothèque	29	7.1762	0.4615	d.f. 1	2
		caution	5	7.3700	0.4684	d.f. 2	11
		non	16	7.5319	0.5524	p-value	0.141856
		All	50	7.3094	0.5099		

The 'Components' panel at the bottom shows various statistical tests, with 'Welch ANOVA' highlighted in the 'Association' category. A red dashed arrow points from the 'View' button in the 'Analysis' tree to the 'Results' tab.

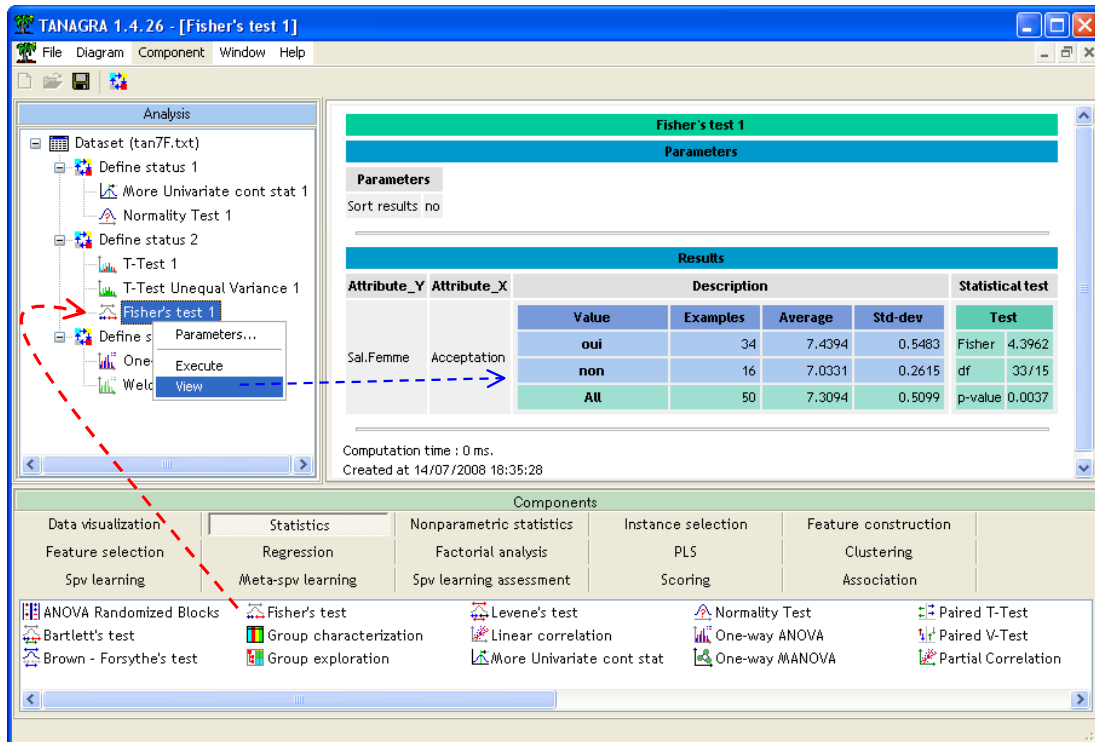
The F-statistic is 2.3446. The first degree of freedom is the number of groups minus 1. The second degree of freedom is more complicated to obtain. We round it to the nearest integer ( $d.f.2 \approx 11$ ). The p-value of the test is 0.141856. The deviation between the means is not significant at the 5% level.

#### 4.5 Comparison of 2 variances – F-Test

The tests for differences in variances are often used prior to the tests for differences in means. They would thus determine the most appropriate procedure. But their utilization is not limited to this task. The comparison of the variance can be the purpose of the analysis. For instance, we want to know if the disposal of the students in a classroom influences the disparity of their scores. In this context, the comparison of variances is at least as interesting as the comparison of means.

In our dataset, we want to compare the variances of the female wages according to the acceptance of the credit. We insert the FISHER'S TEST component (STATISTICS tab) into the diagram<sup>10</sup>, under the DEFINE STATUS 2 component. Thus, the target attribute is SAL.FEMME, and the input one is ACCEPTATION. We click on the VIEW menu.

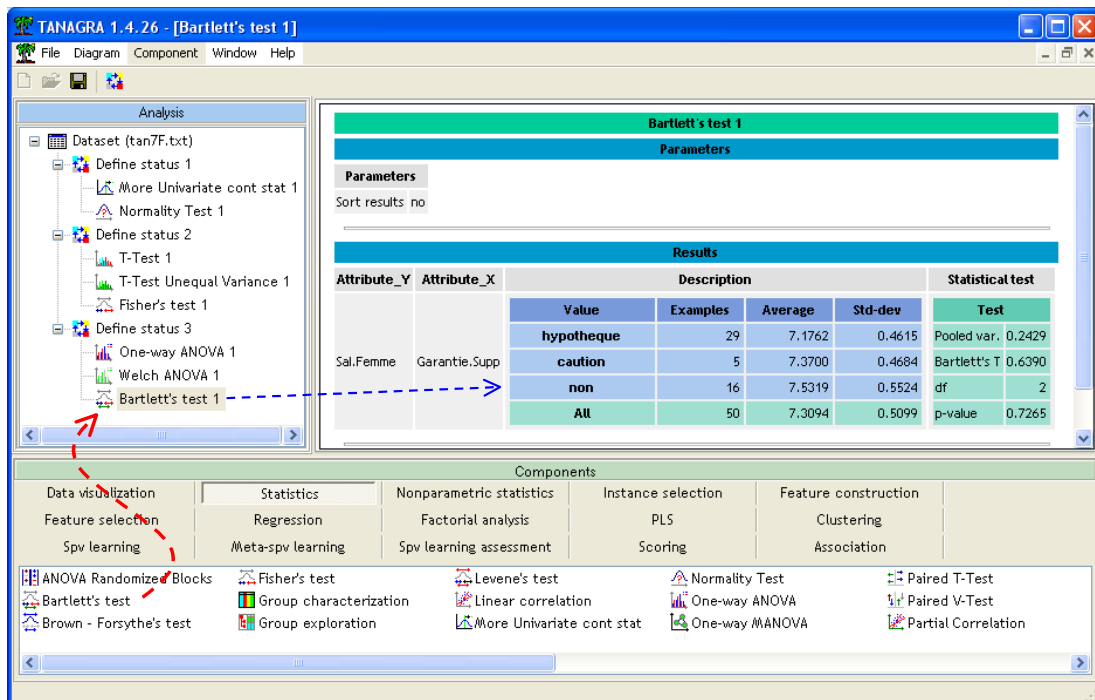
<sup>10</sup> <http://www.itl.nist.gov/div898/handbook/eda/section3/eda359.htm>



The F-statistic is  $F = 4.3962$ . The degrees of freedom are 33 and 15. The p-value of the test is 0.0037. The deviation between the conditional variances is significant. About the comparison of means above, the test for unequal variance seems the most suitable.

#### 4.6 Comparison of K variances – Bartlett's test

The generalization for the comparison of K ( $K \geq 2$ ) variances is the Bartlett's test. We want to compare the variances of the female salary according to the guarantee. We insert the BARTLETT'S TEST (STATISTICS tab) into the diagram.



The Bartlett's statistic is  $T = 0.6390$ , the p-value is  $0.7265$ . The differences in variances are not significant at the 5% level.

#### 4.7 Tests for difference in variances – More robust tests

The F-Test and the Bartlett's test are very sensitive to the departures to the normality assumption<sup>11</sup>. It is more suitable to use a more robust procedure such as the Levene test for equality of ( $K \geq 2$ ) variances<sup>12</sup>. It is reliable even if we deal with a skewed distribution. It is as powerful as the Bartlett's test when we deal with a Gaussian distribution.

##### 4.7.1 The Levene test

We insert the LEVENE'S TEST component (STATISTICS tab) into the diagram.

The screenshot displays the TANAGRA 1.4.26 interface. The 'Analysis' tree on the left shows the workflow including 'Levene's test 1'. The main window shows the 'Levene's test 1' configuration and results. The 'Parameters' section is set to 'Sort results no'. The 'Results' table provides the following data:

Attribute_Y	Attribute_X	Description				Statistical test	
		Value	Examples	Average	Std-dev	Test	
Sal.Femme	Garantie.Supp	hypothèque	29	7.1762	0.4615	Levene's W	0.321175
		caution	5	7.3700	0.4684		
		non	16	7.5319	0.5524		
		All	50	7.3094	0.5099		
						df	2/47
						p-value	0.726876

The 'Components' panel at the bottom shows the 'Statistics' tab selected, with 'Levene's test' highlighted. Other components visible include ANOVA Randomized Blocks, Bartlett's test, Brown - Forsythe's test, Fisher's test, Group characterization, Group exploration, Linear correlation, More Univariate cont stat, Normality Test, One-way ANOVA, One-way MANOVA, Paired T-Test, Paired V-Test, and Partial Correlation.

The Levene's statistic is  $W = 0.321175$ . Under the null hypothesis (the variances are equals), it follows a Fisher distribution with  $(2 ; 47)$  degrees of freedom. The p-value of the test is  $p\text{-value} = 0.726876$ . The differences in variances are not significant at the 5% level.

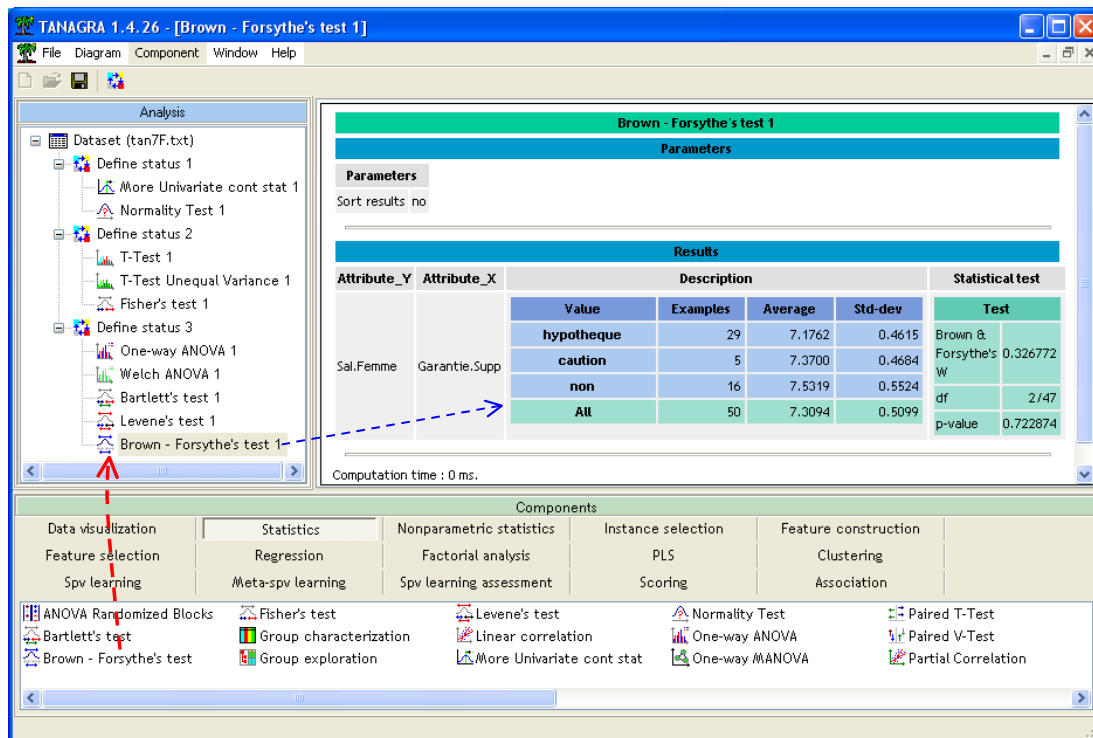
##### 4.7.2 The Brown-Forsythe's test

This version seems the most robust among the tests for equality of variances<sup>13</sup>. We insert the BROWN-FORSYTHE'S TEST component (STATISTICS tab) into the diagram.

<sup>11</sup> <http://www.itl.nist.gov/div898/handbook/eda/section3/eda357.htm>

<sup>12</sup> <http://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm>

<sup>13</sup> [http://en.wikipedia.org/wiki/Brown-Forsythe\\_test](http://en.wikipedia.org/wiki/Brown-Forsythe_test)



## 5 Tests for dependent samples

The aim of the utilization of the dependent samples is to reduce (to control) the uncertainty due to within group variation. It enables to increase the sensitivity of the test<sup>14</sup>.

We say also matched-pairs samples or paired samples.

### 5.1 Paired sample t-test<sup>15</sup>

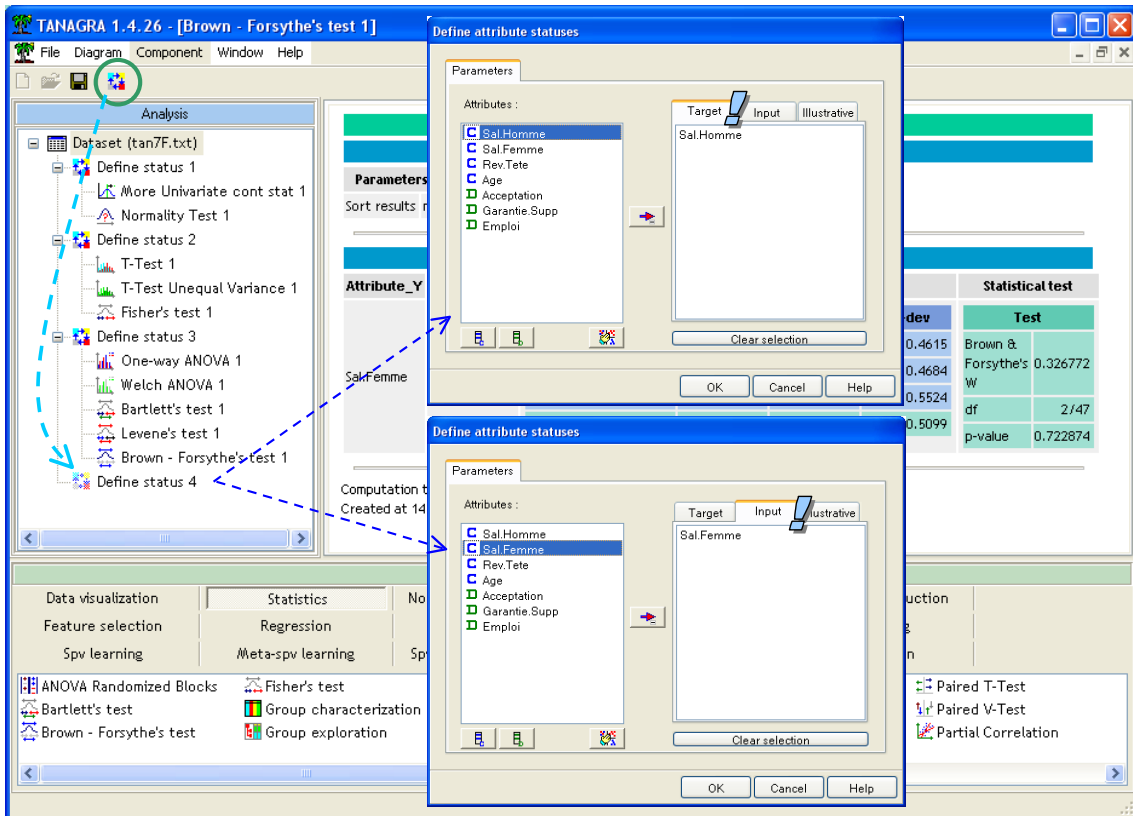
We want to know if in a household, the man tends to have a different salary than his wife.

It is important not to implement a test for independent samples i.e. compare the mean of wages for men with the mean for women. Indeed, **the comparison must be done within each household**. We are in a test scheme for paired samples.

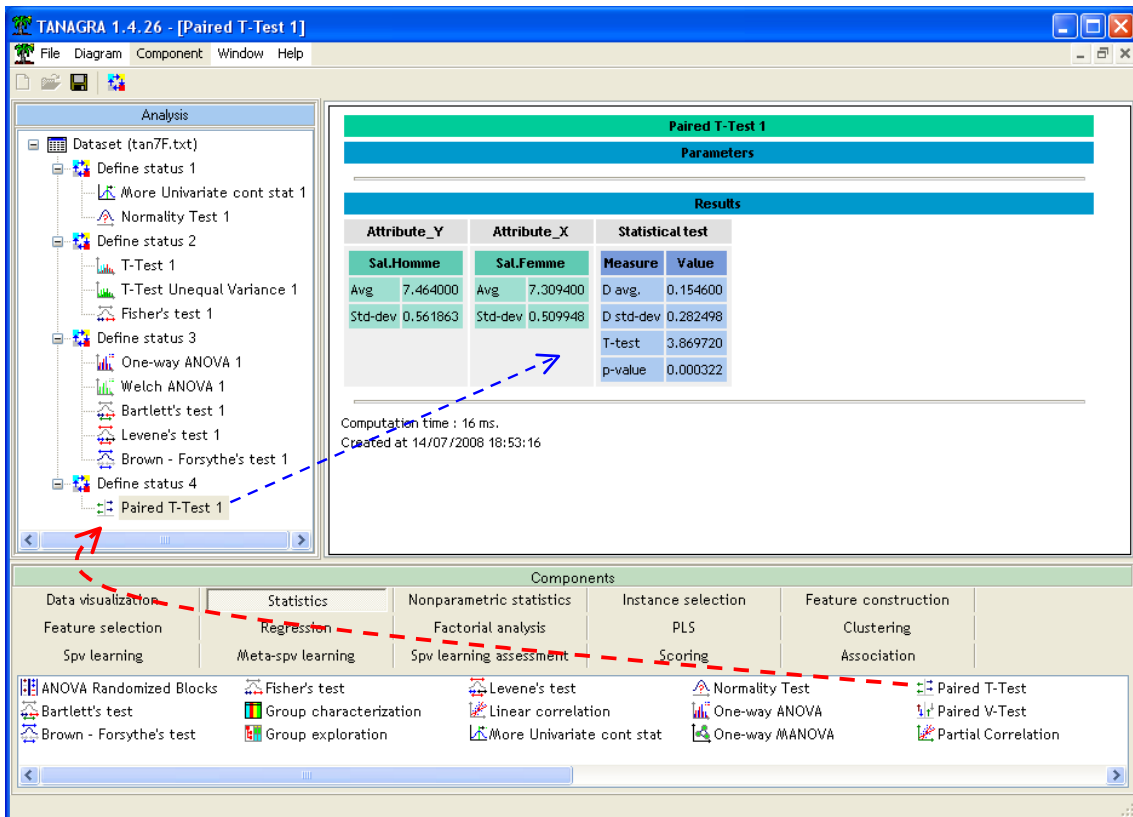
We insert again the DEFINE STATUS component into the diagram. We set SAL.HOMME as target, SAL.FEMME as INPUT.

<sup>14</sup> <http://www.jerrydallal.com/LHSP/paired.htm>

<sup>15</sup> [http://www.stat.psu.edu/online/development/stat500/lesson10/lesson10\\_o4.html](http://www.stat.psu.edu/online/development/stat500/lesson10/lesson10_o4.html)  
[http://www.statsdirect.com/help/parametric\\_methods/ptt.htm](http://www.statsdirect.com/help/parametric_methods/ptt.htm)  
[http://en.wikipedia.org/wiki/Student%27s\\_t-test#Dependent\\_t-test\\_for\\_paired\\_samples](http://en.wikipedia.org/wiki/Student%27s_t-test#Dependent_t-test_for_paired_samples)



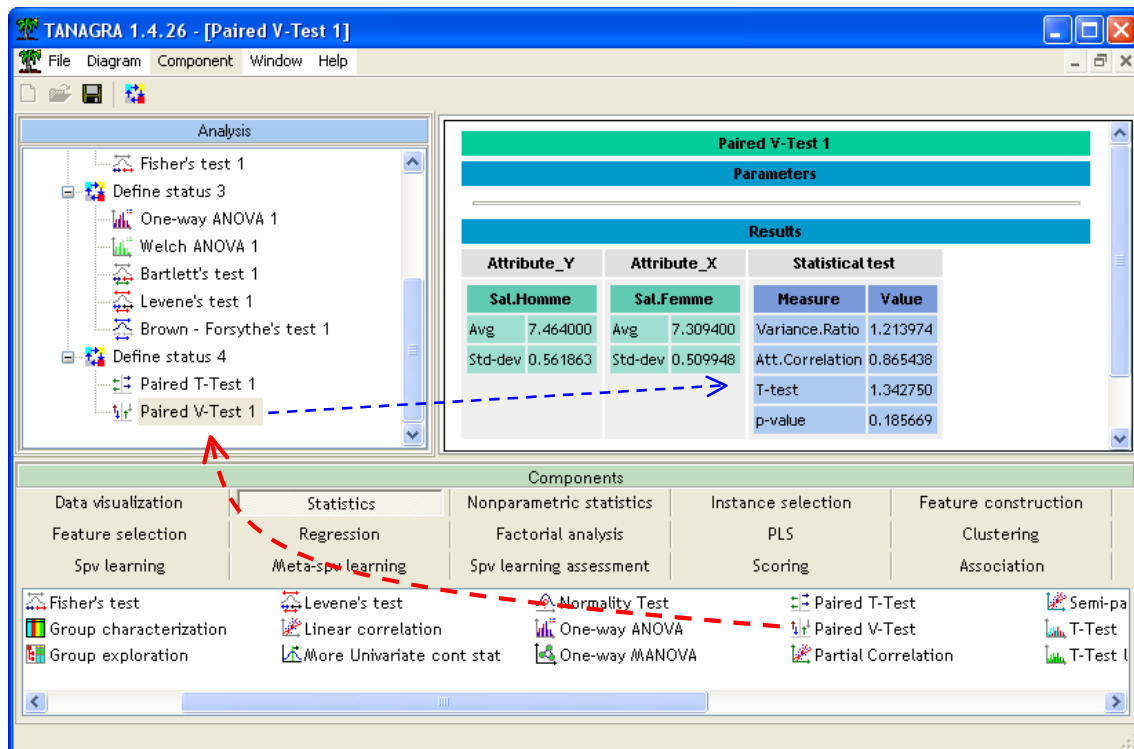
We add the PAIRED T-TEST component (STATISTICS tab) into the diagram.



The mean of the difference between the male and female salaries within households is D.AVG = 0.1546. The test statistic is  $t = 3.86972$ , with a p-value = 0.000322. The difference is significant.

## 5.2 Comparison of variances for 2 paired samples

We can apply the same approach for the comparison of variances. We insert the PAIRED V-TEST component (STATISTICS tab) into the diagram.



The standard deviation of the male salary is 0.561863; for the female, it is 0.509948. The ratio between the variances is  $(0.561863)^2 / (0.509948)^2 = 1.213974$ . But the male and female salaries are highly correlated with  $r = 0.865438$ . The test statistic is then  $t = 1.342750$ . The p-value of the test is 0.185669. The difference between the variances is not significant at the 5% level.

## 5.3 Comparison of means for K paired samples

We can extend the comparison of means for two dependent samples to the comparison for  $K > 2$  paired samples. In the design of experiments, we say "randomized complete block design"<sup>16,17,18</sup>.

We want to implement the ANOVA RANDOMIZED BLOCKS (STATISTICS tab) on the comparison of wages (section 5.1). We can thus compare the results of the two approaches.

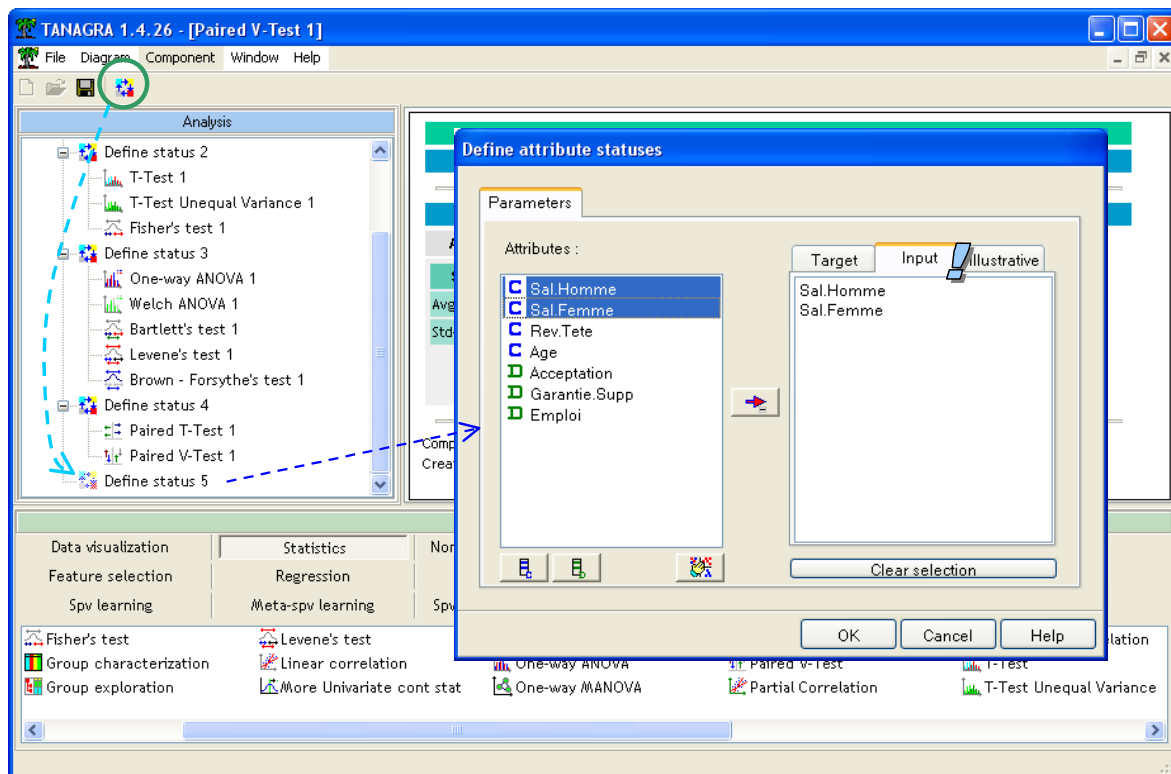
The definition of the status of the variable is a little bit different for this component. We add the DEFINE STATUS component. We set both SAL.HOMME and SAL.FEMME as input.

**Note:** We can set as many variables as we want in INPUT. Simply they must be continuous variables.

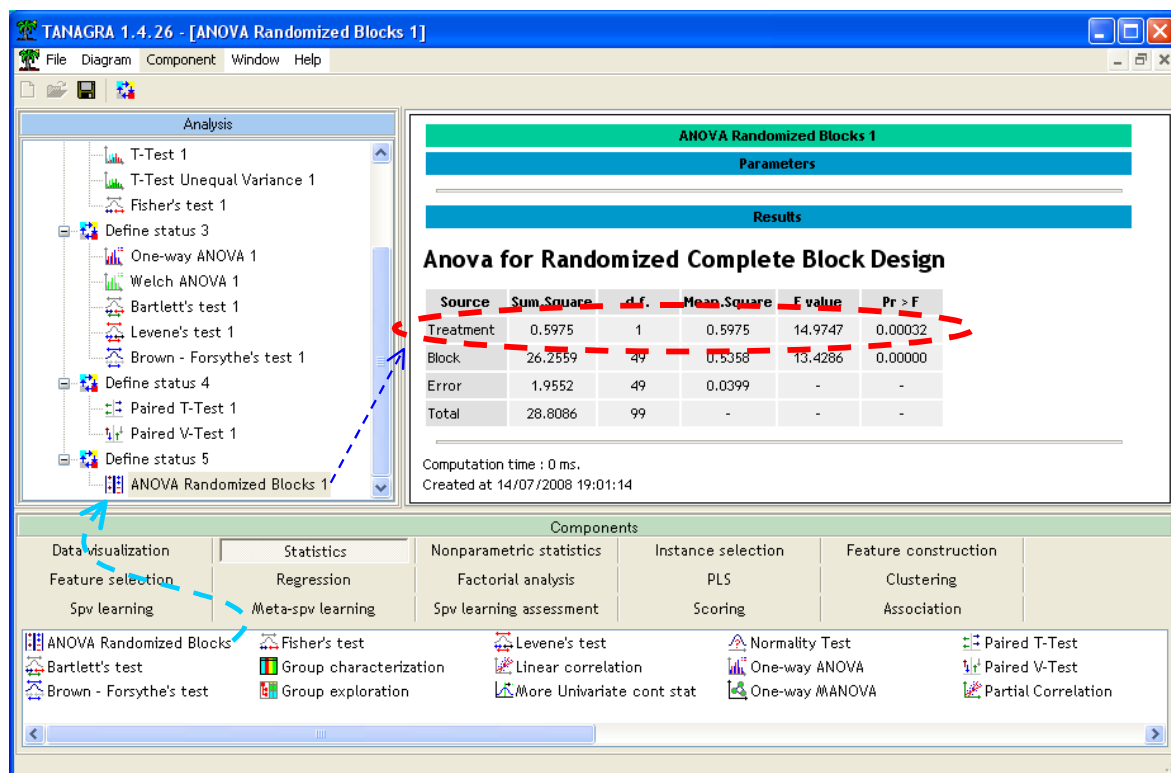
<sup>16</sup> [http://en.wikipedia.org/wiki/Randomized\\_block\\_design](http://en.wikipedia.org/wiki/Randomized_block_design)

<sup>17</sup> <http://itl.nist.gov/div898/handbook/pri/section3/pri332.htm>

<sup>18</sup> <http://www.jerrydallal.com/LHSP/ranblock.htm>



We insert the ANOVA RANDOMIZED BLOCKS component into the diagram.



What interests us is the primary source of variability explained by the "treatment" i.e. the difference between SAL.HOMME and SAL.FEMME. The test statistic  $F = 14.9747$  follows a Fisher distribution at (1, 49) degrees of freedom under the null hypothesis. Again, the difference is significant with a p-value of 0.00032.

**Equivalence with the paired t-test for K = 2 populations.** We know that there is a relation between the Student (T) and the Fisher (F) distributions. Indeed,  $[T(m)]^2 \equiv F(1,m)$ . About our tests, we observe that  $\sqrt{F} = \sqrt{17.9747} = 3.86972 = t$ . It is the value of t-statistic for the paired t-test (section 5.1). The approaches are totally equivalent for K = 2.