1 Subject

Multivariate parametric hypothesis testing for comparing populations.

A multivariate test for comparison of population try to determine if K ($K \geq 2$) samples come from the same underlying population according to a set of variables of interest (X1,...,Xp).

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 data is drawn from a multivariate Gaussian distribution, the hypothesis testing relies on mean vector or on covariance matrix.

The test that we describe in this tutorial can be used to compare process (e.g. is that two machines produce bolts with the same diameter and thickness?), 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 and less fuel consumption on average than men on the roads).

2 Dataset

The CREDIT_APPROVAL.XLS[1] 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 variables of interest; the categorical variables are used to define the sub-populations.

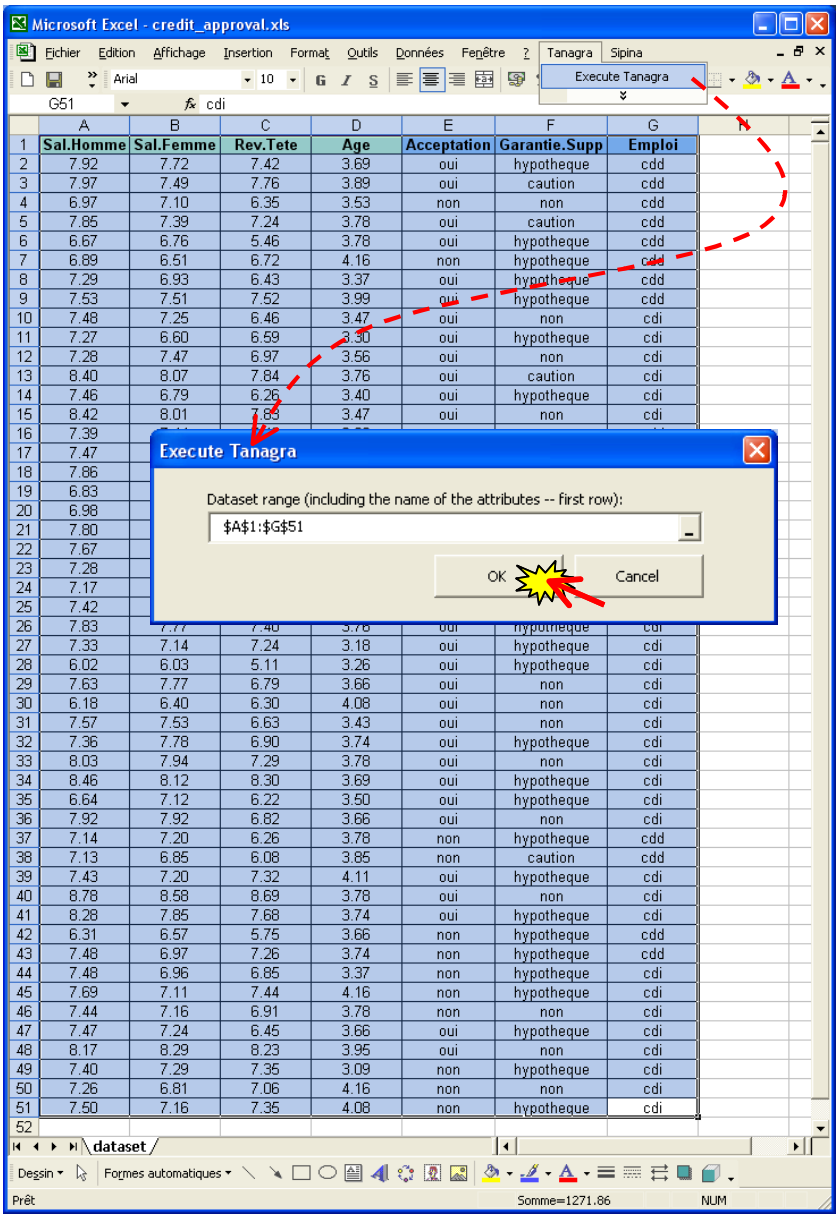
3 Comparison of 2 means – Hotelling's T^2

It is the multivariate analog of the t-test for two independent samples. The size of the mean vector is (p x 1), where p is the number of dependent variables (or variables of interest). The j^{th} component of the vector corresponds to the mean of the variable Xj. We have p = 4 for our dataset.

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

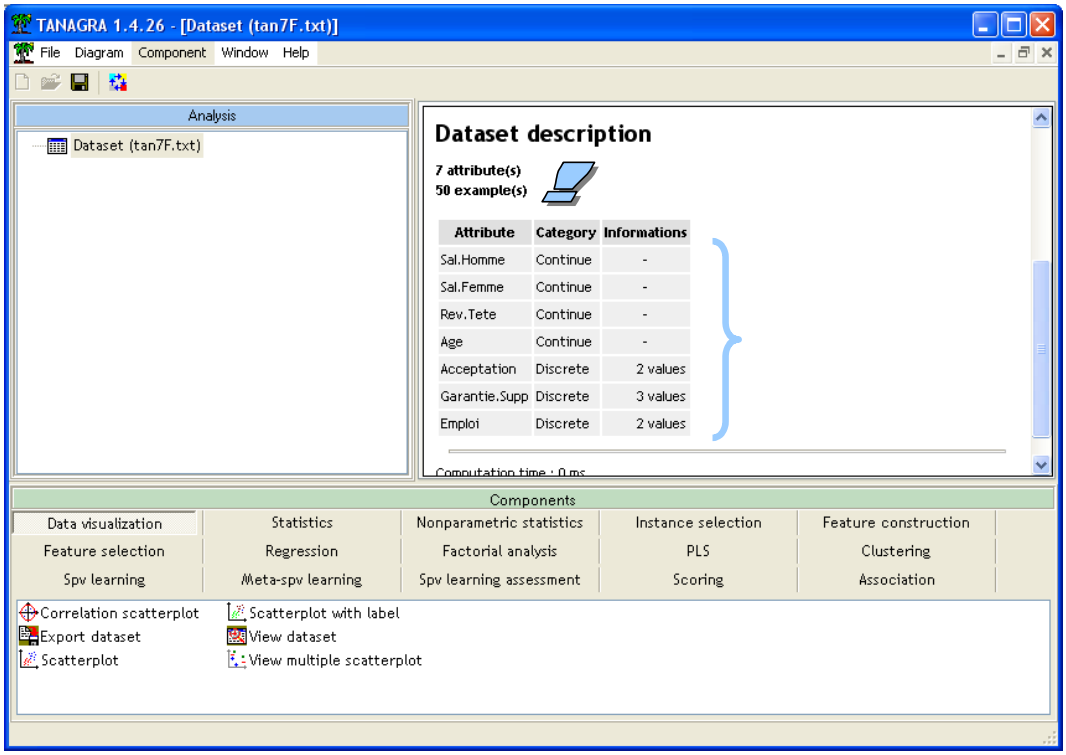
3.1 Importing the dataset

We open the data file into the Excel spreadsheet. We select the data range, and then we click on the Tanagra menu which is installed with the Tanagra.xla add-in[2]. A dialog box appears, we check the coordinate of the cells and we validate by clicking on OK.



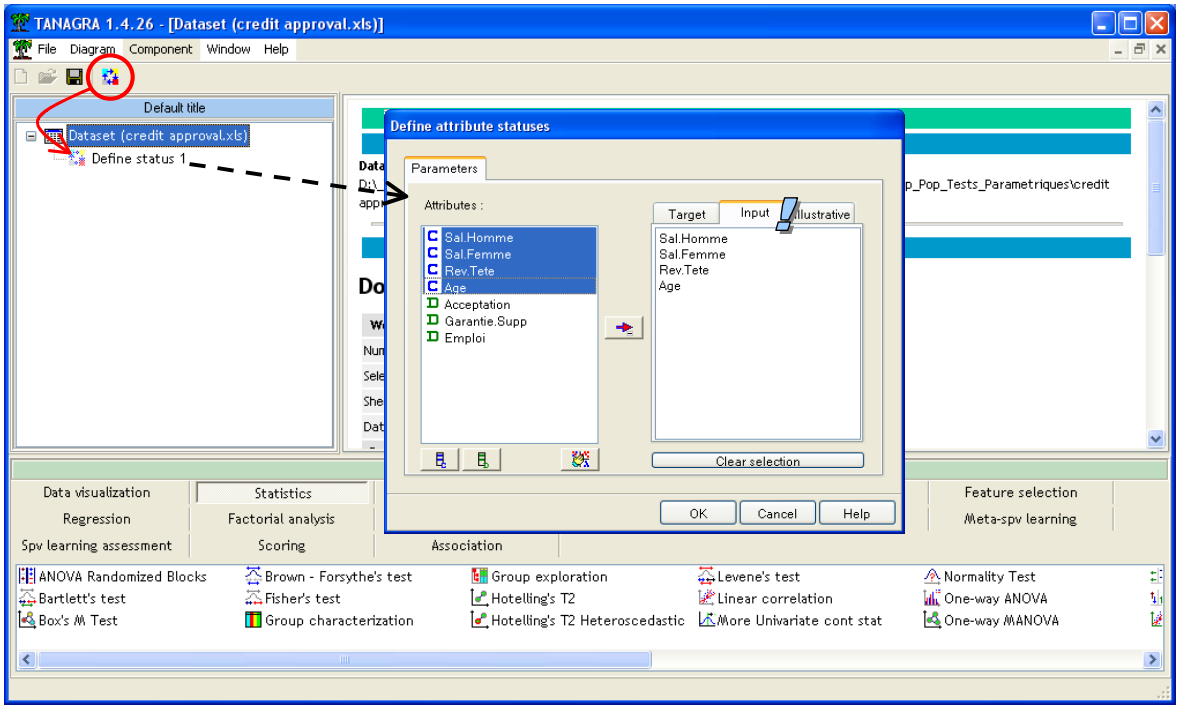
Tanagra is automatically launched. A new diagram is created. We check that the data set has 50 observations and 7 variables.

[2] See http://data-mining-tutorials.blogspot.com/2008/10/excel-file-handling-using-add-in.html
Tanagra can also directly read a XLS file format, see http://data-mining-tutorials.blogspot.com/2008/10/excel-file-format-direct-importation.html

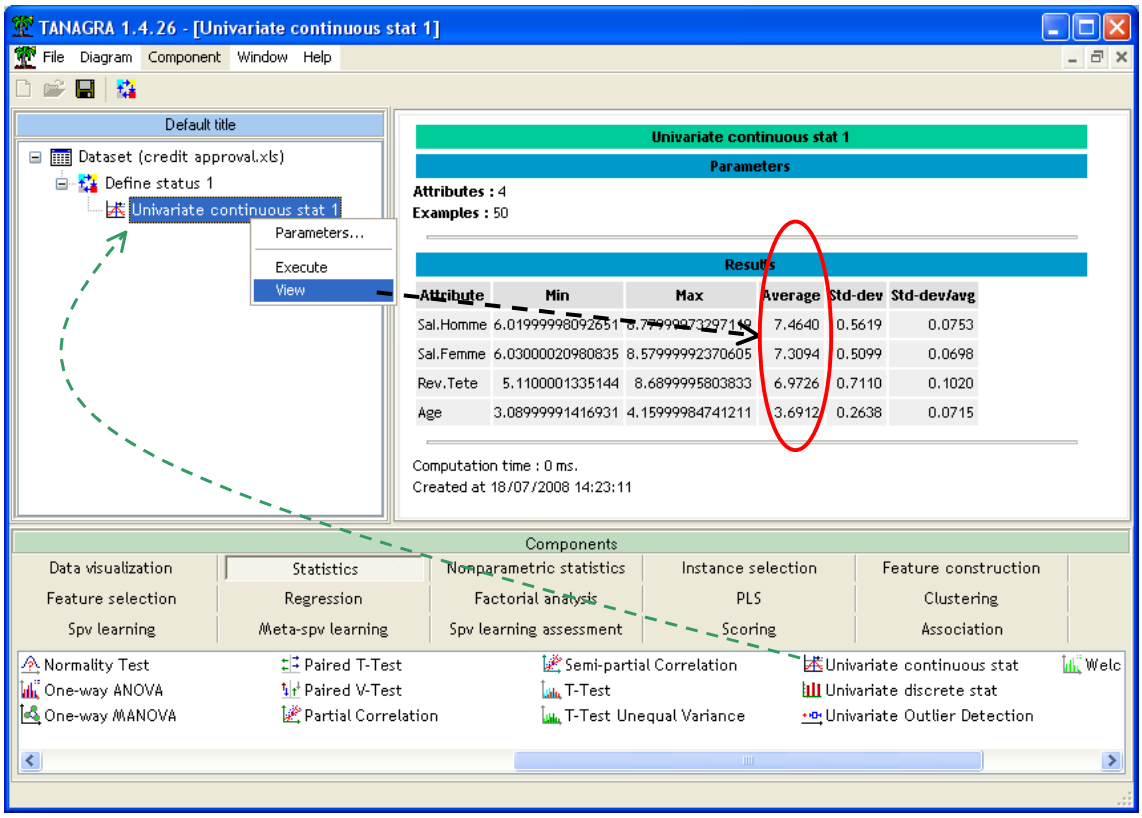


3.2 Descriptive statistics

First, we want to describe the dependent variables. We insert the DEFINE STATUS component in the diagram using the shortcut into the toolbar. We set all the continuous variables as INPUT.



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



The AVERAGE column is particularly interesting in this step. We read the mean vector computed on the available observations $\overline{X}' = (7.4640; 7.3094; 6.9726; 3.6912)$. The aim of the Hotelling's T2 is to verify that this vector takes values significantly different in the subpopulations that we wish to compare.

3.3 Hotelling's T² under the homoscedasticity context

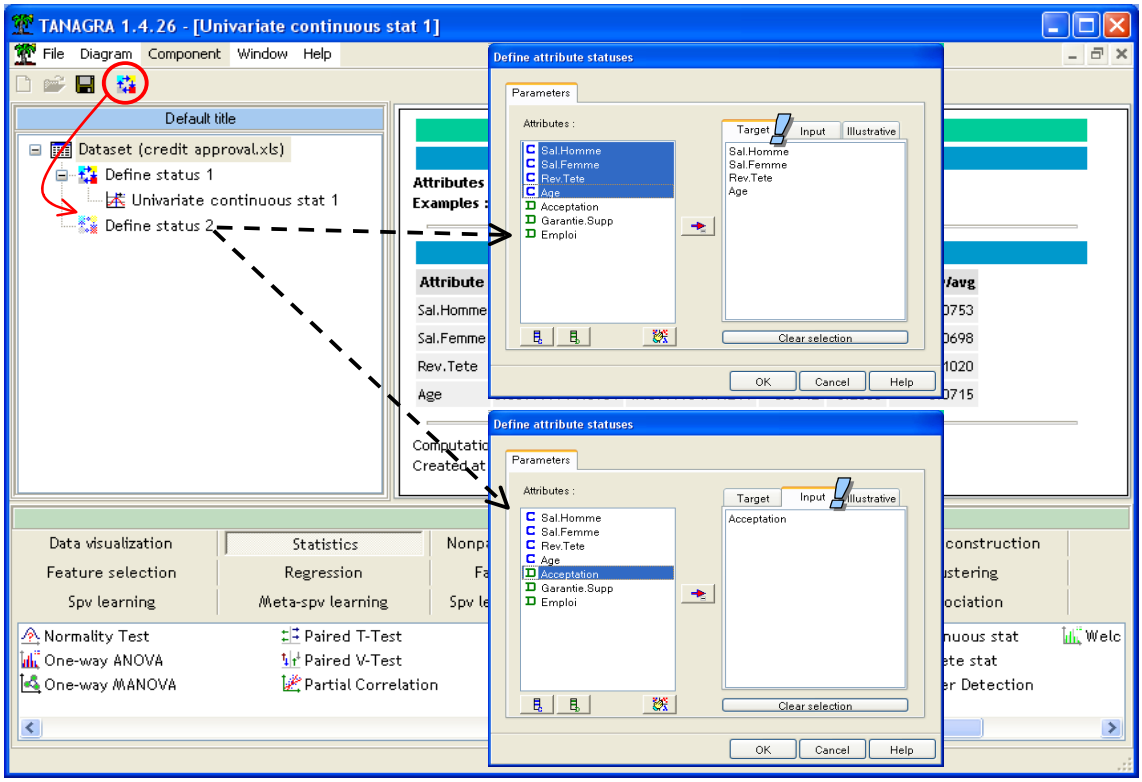
We want to compare two mean vectors. We make the assumption that the covariance matrices are identical into the two subpopulations[3]. We thus used the pooled covariance matrix[4] into the formula.

On our dataset, we want to compare the characteristics of the clients according the acceptation of the reject of the credit.

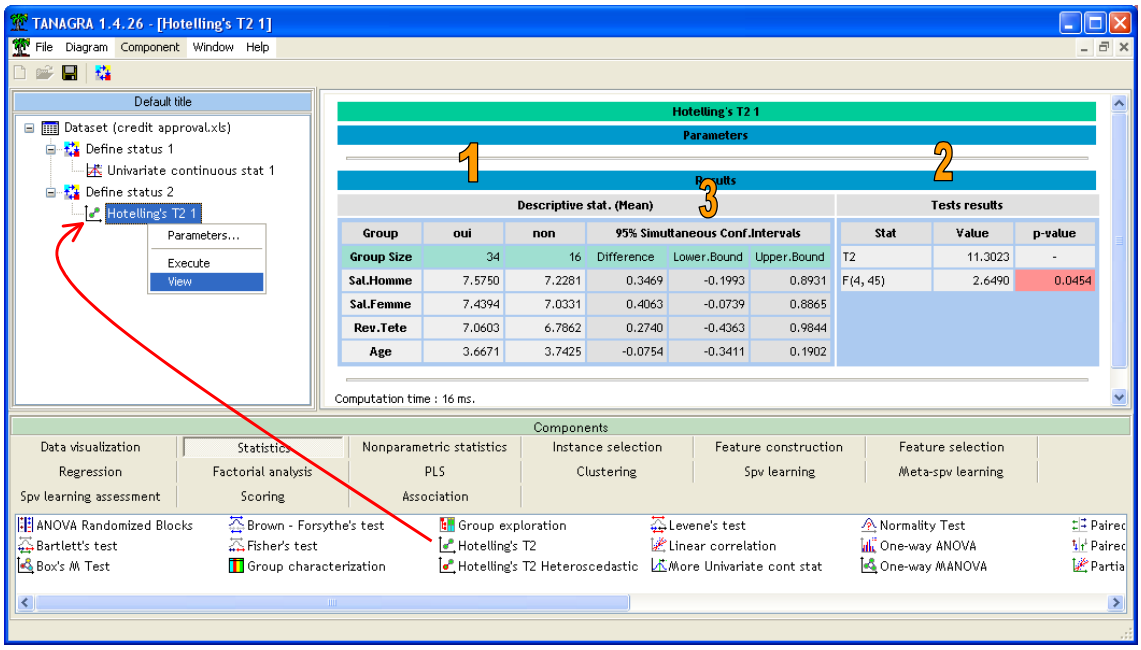
We insert the DEFINE STATUS component into the diagram. We set the dependent variables as TARGET, credit ACCEPTATION as INPUT.

[3] http://en.wikipedia.org/wiki/Homoscedasticity

[4] http://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/poolcov.htm



We add the HOTELLING'S T2 component into the diagram. We click on the VIEW menu.



Several results draw our attention:

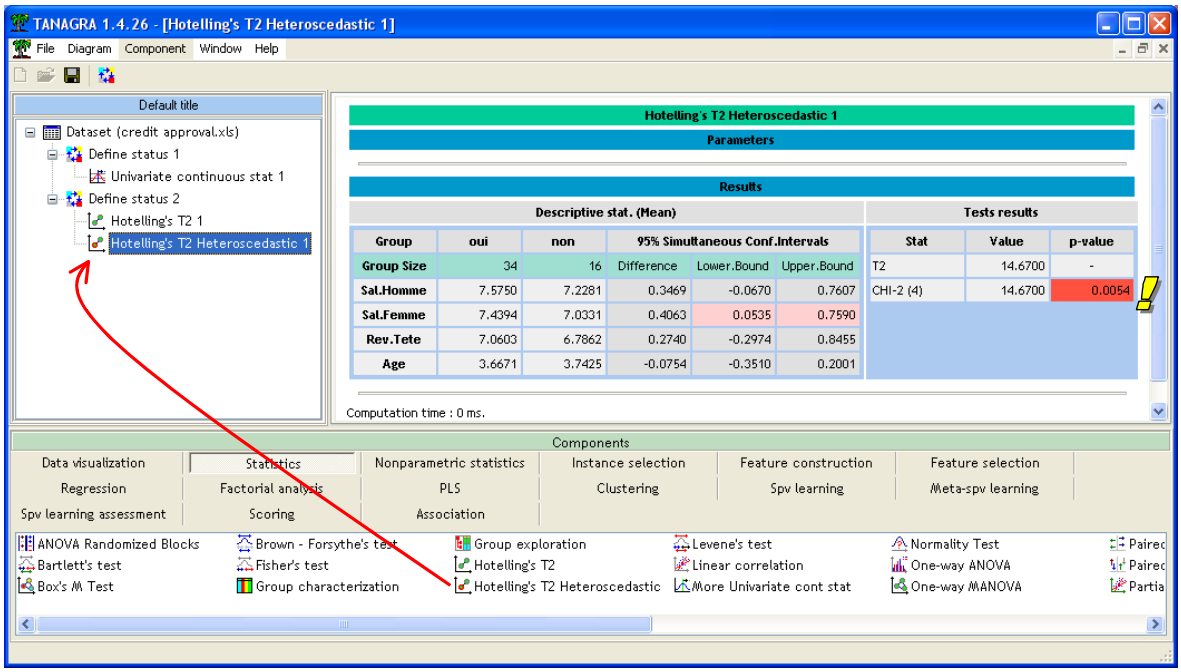
- (1) Into the descriptive statistics table, we observe the mean vector for each group (acceptation = OUI and acceptation = NON).
- (2) In the right, we have the T2 statistic, and its transformation which is distributed as a Fisher distribution with (4; 45) degrees of freedom. The p-value of the test is 0.0454, we reject the homogeneity of the mean vectors at the 5% significance level.

- (3) Let's we consider again the descriptive statistics table. Into the 3 last columns, we have the vector of differences between means and their simultaneous confidence intervals at 95% confidence level. These intervals allow us to know if the difference is significant at 5% on one of the dependent variables that we want to analyze in particular. The difference is significant if the interval does not contain the value 0. In our example, it seems that the overall significant difference detected by the T^2 is not specifically associated with one of the dependent variables.

3.4 Hotelling's T^2 under the heteroscedasticity context

If the covariance matrices are different, especially when the size of groups is unequal, we must compute separately the conditional covariance matrices. The test statistic is thus different.

Under the DEFINE STATUS into the diagram, we add the HOTELLING'S T2 HETEROSCEDASTIC (STATISTICS tab).

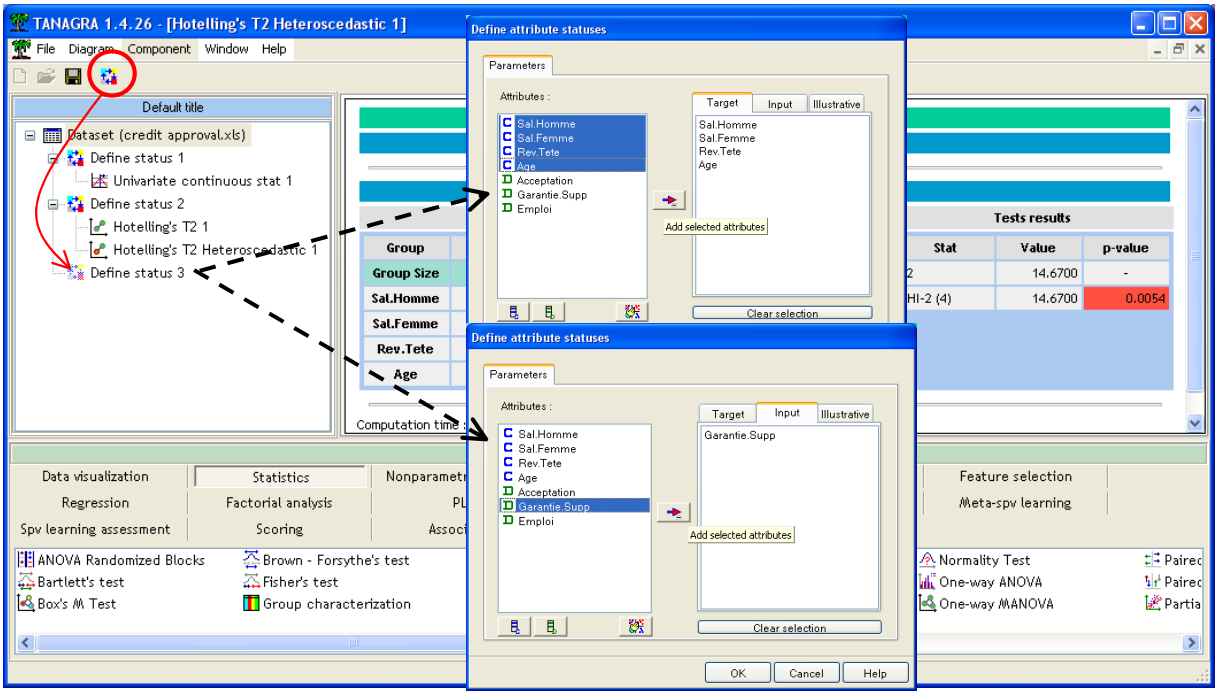


The T^2 statistic is equal to $T^2 = 14.6700$, it is distributed as the CHI-square distribution with p degrees of freedom. The deviation between the mean vectors is significant at the 5% level (p-value = 0.0054). It seems now that this deviation relies mainly on the differences between the wages of the wife into the household. The confidence interval does not contain the value 0 at 95% confidence level.

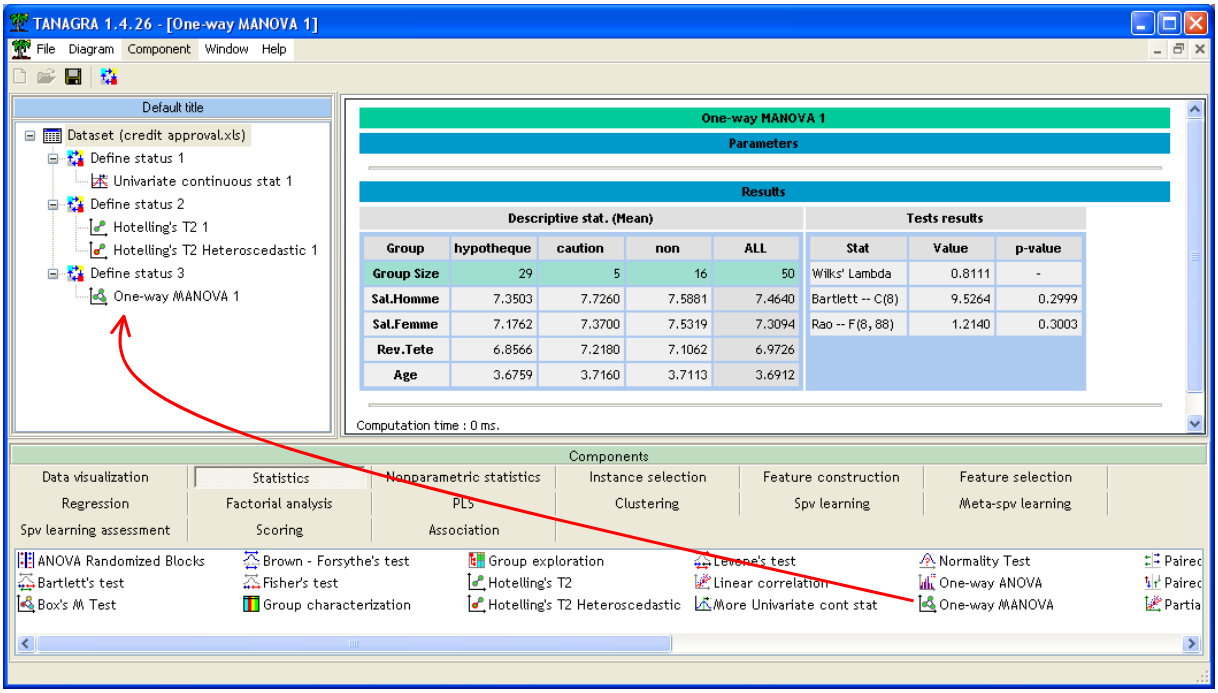
4 Comparing mean vector for K populations (MANOVA)

We want to compare the mean vectors for K ($K \geq 2$) populations. It is the multivariate analog of the one-way analysis of variance (ANOVA). It is often called MANOVA (multivariate analysis of variance). On our dataset, we want to compare the mean vectors according to the kind of guarantee ($K = 3$ values) requested by the debtors.

We add the DEFINE STATUS component into the diagram. We set as TARGET our dependent variables, and we set GARANTIE.SUPP as INPUT.



We add the ONE-WAY MANOVA component (STATISTICS tab).



Into the DESCRIPTIVE STAT part, we observe the conditional mean vectors for the 3 groups, and the overall mean. The Wilks' lambda statistic is 0.8111. The more it is close to 1, the less significant are the differences. Two transformations are available. The Bartlett's transformation follows as a chi-square distribution, it is sufficient for large samples. The Rao's transformation is preferable on small

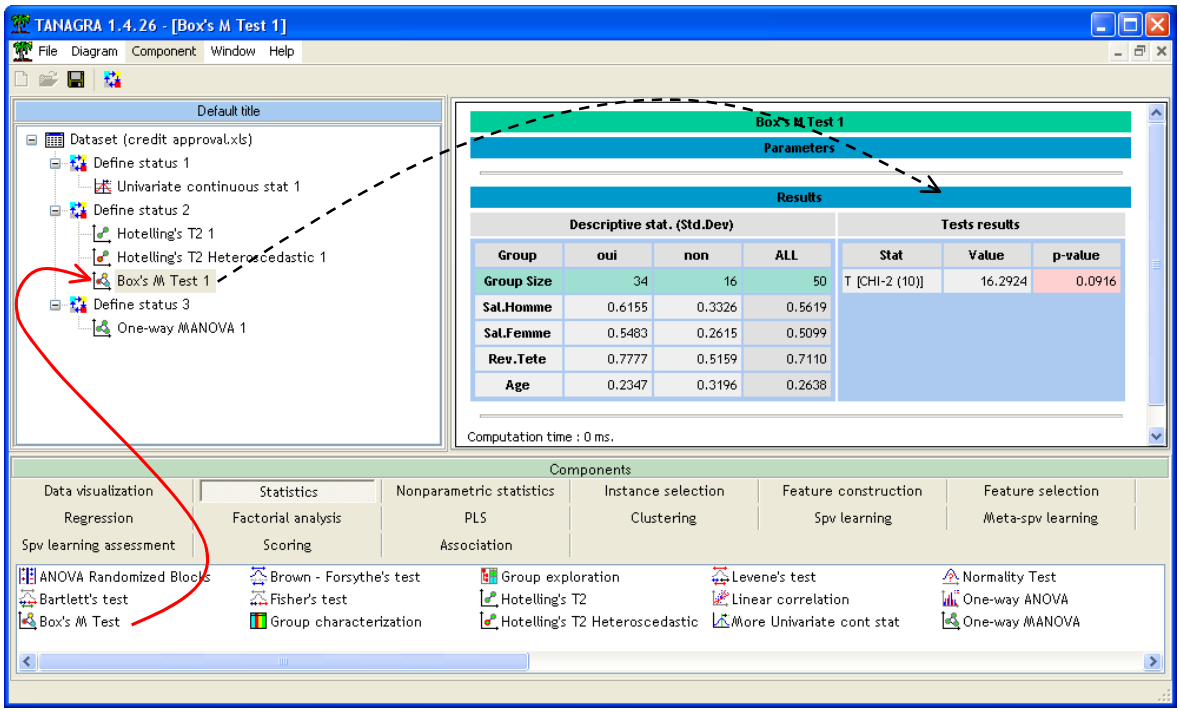
samples. We have the same conclusion on our dataset. The null hypothesis cannot be rejected from the available dataset.

5 Comparing K variances – The Box's M statistic

The first Hotelling's test and the Manova test make the assumption that the conditional covariance matrices are identical. We can check this assumption by using the multivariate analog of the Bartlett's test for the equality of variances. It is called Box's M test. We implement this test on our dataset.

5.1 Checking for the Hotelling's test

We insert the BOX'S M TEST (STATISTICS tab) into the diagram, at the same level as the Hotelling's test component. We click on the VIEW menu in order to obtain the results.



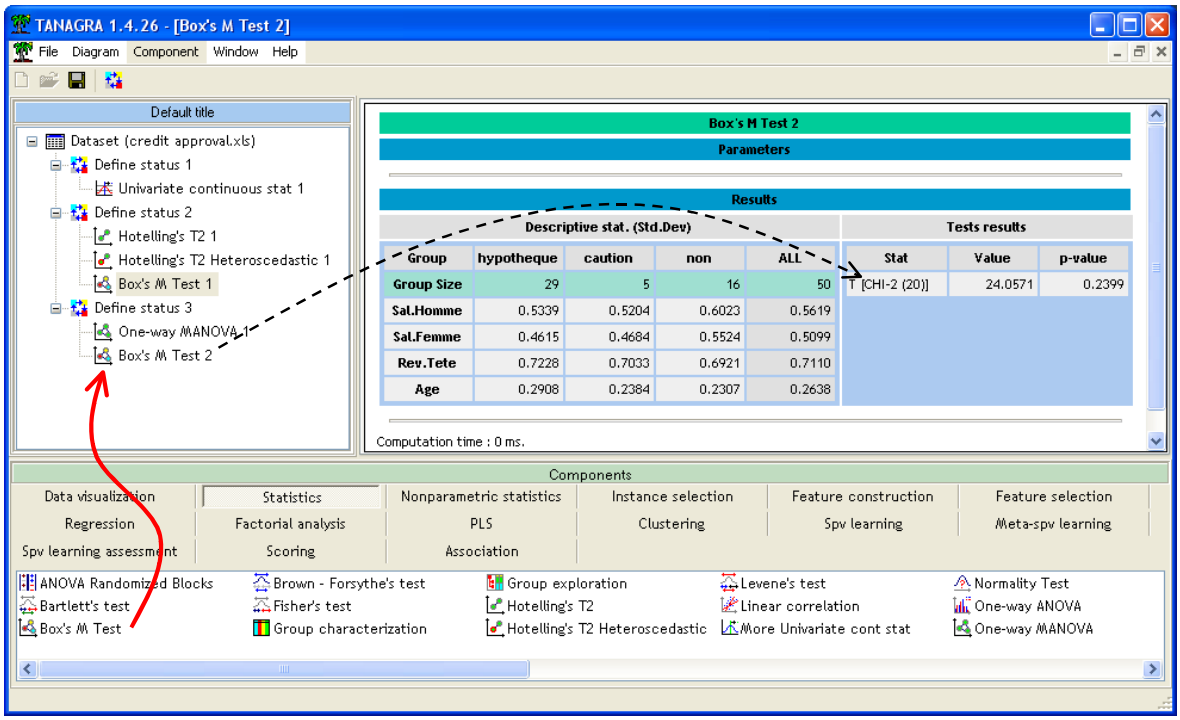
The overall and the conditional standard deviation are supplied into the descriptive statistics table. These values are purely indicative. They give an idea of differences, in univariate way, for each dependent variable. We cannot draw conclusions from these descriptive statistics because they do not take into account of the covariance between the variables.

The test statistic is T = 16.2924. It follows a chi-square distribution with 10 degree of freedom. We conclude that the conditional covariance matrices are not significantly different at 5% level (p-value = 0.0916), but they are at 10% level.

In this context, the homoscedaticity assumption is not obvious, and we have small groups with different size, it is more suitable to use the heteroscedastic version of the Hotelling's test.

5.2 Checking for the Wilks' test

We make the same verification for the MANOVA. We add again the BOX'S M TEST component.



Here, the homoscedasticity assumption is really believable. We have T = 24.0571, with a p-value = 0.2399. It seems that the results associated with Wilks' Lambda when comparing the mean vectors of the K groups (according to the guarantee) are confirmed.