

1. Subject

Understanding the “test value” criterion (“Valeur test” in French, VT).

The test value (VT) is a criterion often used in various components of TANAGRA. It is mainly used for the characterization of a group of observations according a continuous or categorical variable. The groups may be defined by categories from a discrete variable; they can also be computed by a machine learning algorithm (e.g. clustering, a node of a decision tree, etc.).

The principle is elementary: we compare the values of a descriptive statistic indicator computed on the whole sample and computed on sub sample related to the group. For a continuous variable, we compare the mean; for a discrete one, we compare the proportion.

Despite, or because of its simplicity, the VT is very useful. The formulation that we present in this tutorial is taken from the Lebart et al.'s book (2001)¹. The VT is intensively used in some commercial software such as SPAD (<http://eng.spad.eu/>). It allows to characterize groups, but it can be used also to strengthen the interpretation of the factors extracted from a factorial analysis process.

In this tutorial, we emphasis the formulas used for both categorical and continuous variables. We put them in connection with the results provided by the GROUP CHARACTERIZATION component of TANAGRA.

2. Definition and formulas

The size of the whole dataset is n . The size of the sub sample related to a sub population is n_g .

Of course, $n_g < n$, the sub sample is a part of the whole sample.

Comparison according a continuous variable

X is a continuous variable. The mean computed on the whole sample is μ , the empirical variance is σ^2 ; the mean computed into the group is μ_g .

The test value is then defined as follows (Lebart et al., 2000; p. 181):

¹ L. Lebart, A. Morineau, M. Piron, « Statistique Exploratoire Multidimensionnelle », Dunod, pp.181-184, 2000.

$$t_c = \frac{\mu_g - \mu}{\sqrt{\frac{n - n_g}{n - 1} \times \frac{\sigma^2}{n_g}}}$$

We see in the denominator the standard deviation of the mean in the case of a sampling without replacement of n_g elements among n

The test value t_c can be viewed as a statistical test of comparison of means. But the samples are not independent. The indicator follows asymptotically a Gaussian distribution, then for a 5% significance level, we consider that the difference is significant if the absolute value is greater than 2.

But it's not as simple. If the groups are supplied by a learning algorithm, the variable was used for the differentiation of groups. The differences computed afterwards are artificially relevant. Furthermore, when we handle a large database, all the computed statistical indicators seem relevant.

For these reasons, we should not overly focus on the comparison of the computed VT with a threshold, very difficult to define in practice. It is more important to use the VT as a criterion for the ranking of the variable, in order to distinguish the variables that play an essential role in the interpretation of the groups. It was especially important to detect situations where VT of a variable is significantly different from the others.

Comparison according a categorical (discrete) variable

Y is a discrete variable. We focus on the category j . n_j is the number of instance corresponding to this category in the whole sample; n_{jg} is the number of instance corresponding to ($Y = j$) into the sub sample related to the group; the group size is n_g .

If the sub sample is randomly drawn from the whole sample, the expected number of individuals corresponding to ($Y = j$) into the sub sample would be

$$\pi = \frac{n_g \times n_j}{n}$$

The test value is then (Lebart et al., 2000 ; page 184):

$$t_d = \frac{n_{jg} - \frac{n_g \times n_j}{n}}{\sqrt{\frac{n - n_g}{n - 1} \times \left(1 - \frac{n_j}{n}\right) \times \frac{n_g \times n_j}{n}}}$$

Here again, the criterion is mainly used to highlight the category which characterizes the better the group of observations.

For Tanagra, the results (variable = category) is sorted according to the VT, we will focus primarily on the top and bottom of the tables.

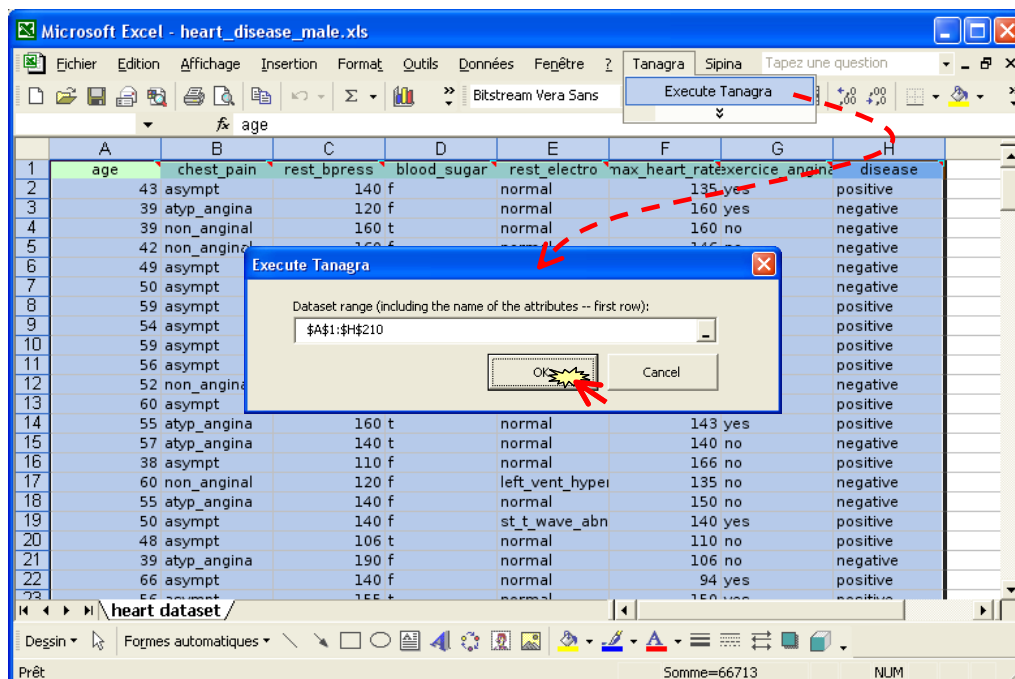
3. Handling and understanding the VT criterion

To illustrate these formulas, we perform a group characterization process with TANAGRA. Unlike the usual tutorials, we will focus, in relation to the above formulas, on the results provided by the software.

Loading the dataset

We use the HEART_DISEASE_MALE.XLS² data file. It describes males which have or not a heart disease. We aim to characterize the individuals which are ill.

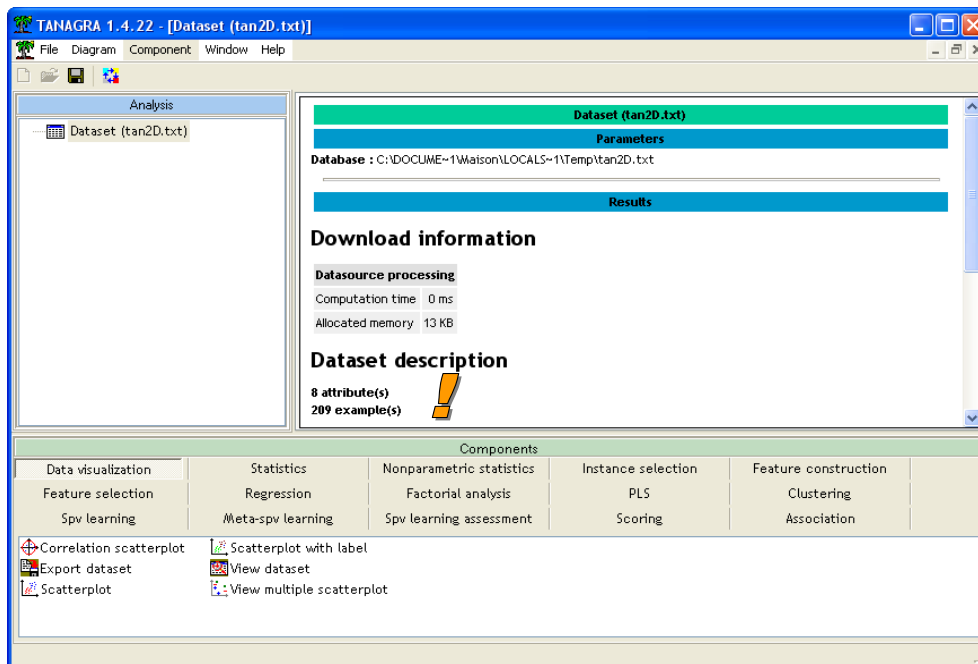
We load the data file into Excel. We select the range of cells. We activate the TANAGRA / EXECUTE TANAGRA menu³. A dialog box appears, we check the selection and then we validate.



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

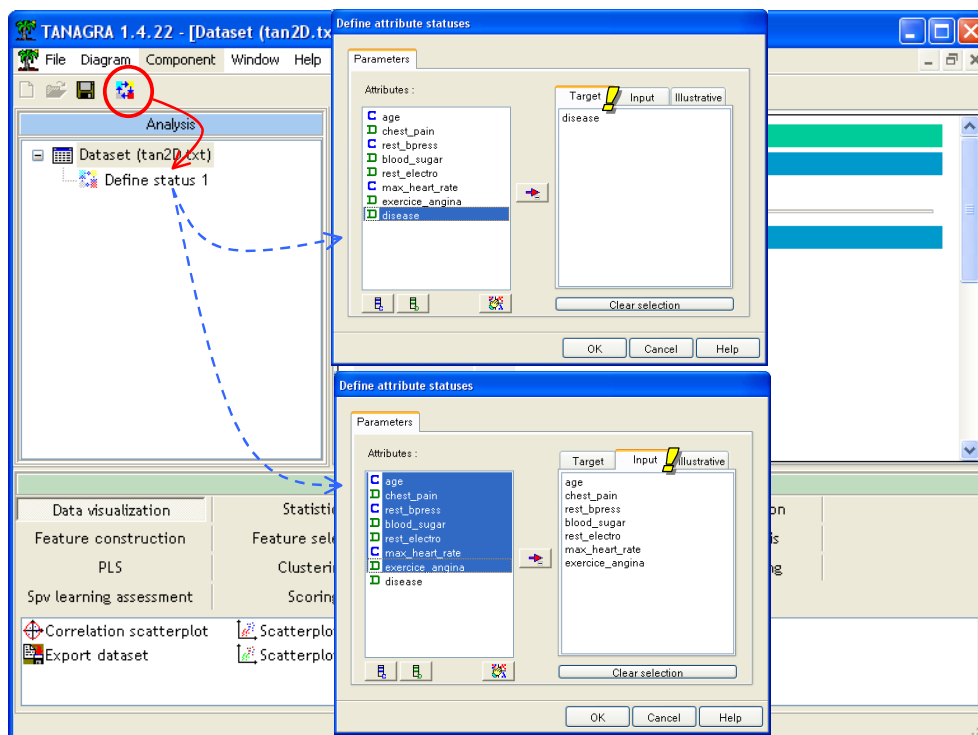
³ See <http://data-mining-tutorials.blogspot.com/2008/10/excel-file-handling-using-add-in.html> ; we can install the add-on also into the Open Office Calc: see <http://data-mining-tutorials.blogspot.com/2008/10/oocalc-file-handling-using-add-in.html> and <http://data-mining-tutorials.blogspot.com/2009/04/launching-tanagra-from-oocalc-under.html>

TANAGRA is automatically launched. A new diagram is created and the data file is loaded. There are $n = 209$ examples into the data file.



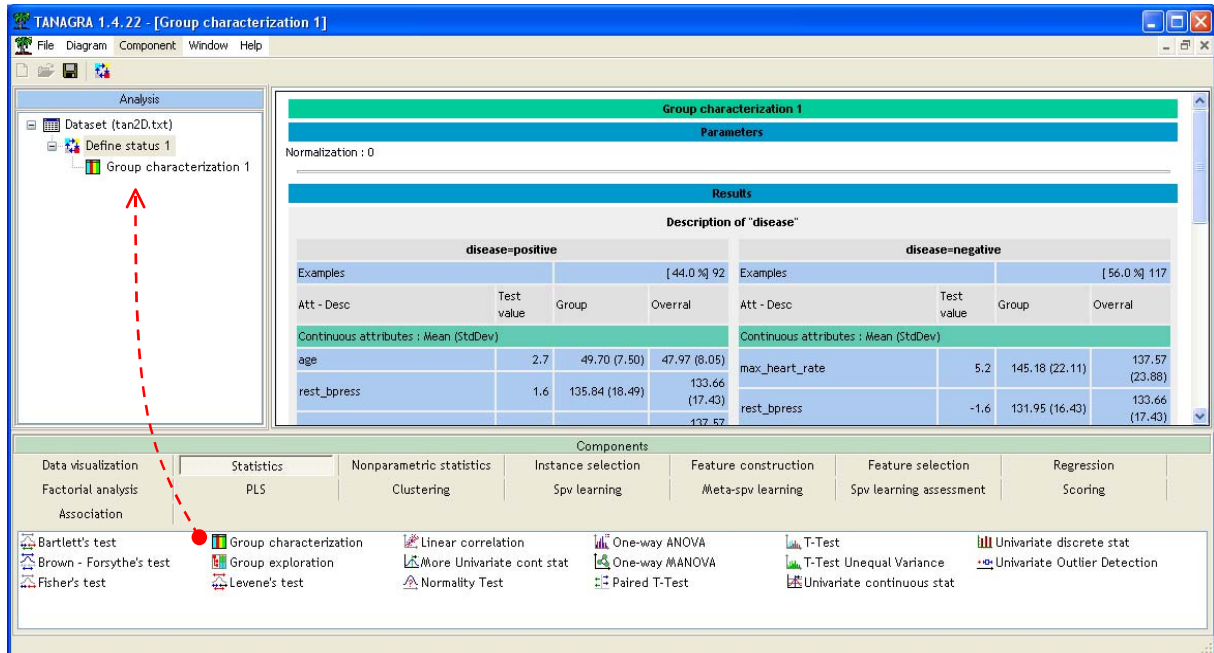
Defining the type of the variables

We want to characterize the group (DISEASE = POSITIVE). We insert the DEFINE STATUS component into the diagram. We set DISEASE as TARGET; the other variables as INPUT.



Characterization of the groups

In order to characterize the observations of the group (DISEASE = POSITIVE), we use the GROUP CHARACTERIZATION component (STATISTICS tab). We click on the VIEW menu.



The size of the sample (DISEASE = POSITIVE) is $n_g = 92$. We analyze in detail the column related to this category.

disease=positive			
Examples	[44.0 %] 92		
Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)			
age	2.7	49.70 (7.50)	47.97 (8.05)
rest_bpress	1.6	135.84 (18.49)	133.66 (17.43)
max_heart_rate	-5.2	127.90 (22.61)	137.57 (23.88)
Discrete attributes : [Recall] Accuracy			
chest_pain=asympt	8.4	[73.5 %] 81.5 %	48.80%
exercice_angina=yes	8.3	[83.3 %] 65.2 %	34.40%
blood_sugar=t	2.1	[68.8 %] 12.0 %	7.70%
rest_electro=st_t_wave_abnormality	1.5	[56.7 %] 18.5 %	14.40%
chest_pain=typ_angina	1.1	[66.7 %] 4.3 %	2.90%
rest_electro=?	1.1	[100.0 %] 1.1 %	0.50%
rest_electro=left_vent_hyper	-1.1	[20.0 %] 1.1 %	2.40%
rest_electro=normal	-1.2	[42.2 %] 79.3 %	82.80%
blood_sugar=f	-2.1	[42.0 %] 88.0 %	92.30%
chest_pain=non_anginal	-3.3	[19.4 %] 7.6 %	17.20%
chest_pain=atyp_angina	-6.8	[9.2 %] 6.5 %	31.10%
exercice_angina=no	-8.3	[23.4 %] 34.8 %	65.60%

Continuous variable

We study the AGE variable. The mean of AGE in the whole population (sample) is 47.97. Into the group (DISEASE = POSITIVE), the mean becomes 49.70. Into the brackets, we have the computed standard deviation. The test value is:

$$t_c = \frac{49.70 - 47.97}{\sqrt{\frac{209 - 92}{209 - 1} \times \frac{8.05^2}{92}}} \approx 2.74$$

We see this result into the table above. In this group, the individuals seem older.

Categorical variable

The first category (ATTRIBUTE = VALUE) which seems characterize the ill individuals is CHEST PAIN = ASYMPT. In the whole population, 48.8% of the individuals have the characteristic CHEST PAIN = ASYMPT; into the group DISEASE = POSITIVE, the proportion becomes 81.5%; furthermore, 73.5% of the individuals corresponding to (CHEST PAIN = ASYMPT) are in this group.

The details of the calculations are: $n = 209$, $n_g = 92$, $n_j = 0.488 \times 209 = 102$,
 $n_{jg} = 0.815 \times 92 = 75$,

Then

$$t_c = \frac{75 - \frac{92 \times 102}{209}}{\sqrt{\frac{209 - 92}{209 - 1} \times \left(1 - \frac{102}{209}\right) \times \frac{92 \times 102}{209}}} \approx 8.37$$

4. Conclusion

The characterization of groups is a key task for the data exploration. The univariate analysis, based on comparisons of means or frequencies, even if it seems simplistic, is often very valuable. The "test value" criterion gives us the opportunity to rank the variables according to their relevance.