Subject

Association measures for ordinal variables.

In this tutorial, we show how to use TANAGRA (1.4.19 and higher) for measuring the association between ordinal variables.

All the measures that we present here rely on the concept of pairs. A good reference about this concept is the following: http://www2.chass.ncsu.edu/garson/PA765/association.htm.

About the utilization and the interpretation of these measures, there is another good reference (http://www2.chass.ncsu.edu/garson/PA765/assocordinal.htm).

The used formulas are available on-line -- http://v8doc.sas.com/sashtml/stat/chap28/sect20.htm.

If, in a theoretical point of view, the measures intended for continuous attributes such as correlation are not convenient in our context, in the practical point of view, we display, in this tutorial, that it nevertheless gives interesting results for the studying the dependence between ordinal variables.

Dataset

The used dataset come from a case study available on the web[1]. The aim is to predict the high blood pressure (hypertension) from the characteristics of patients.

Dependant variable

The original dependent variable is SYSTOLIC. We discretize it into 3 intervals (BP3Levels). We use the usual cut points in order to characterize the degree of hypertension

- Normal if SYSTOLIC <= 140 mm hg
- High if SYSTOLIC > 140 mm g and <= 180 mm hg
- Very high if SYSTOLIC > 180 mm hg

Note: cut points for discretization. As we will see below, we note that the choice of the number of categories of the discretized variable has an influence on the results. We will see that using a two-level blood pressure (BP2Levels), association which does not seem statistically significant with BP3LEvels becomes significant when we used BP2Levels. For the determination of high blood pressure, we use 9 independent variables:

Variables	Description
Gender_M	gender (1 : male ; 0 : female)
Smoke_Y	Smoke (1 : yes ; 0 : no)
Exercise	Exercise level (1 : low ; 2 : medium ; 3 : high)
Overweight	Overweight (1 : normal ; 2 : overweight ; 3 : obese)
Alcohol	Alcohol use (1 : low ; 2 : medium ; 3 : high)

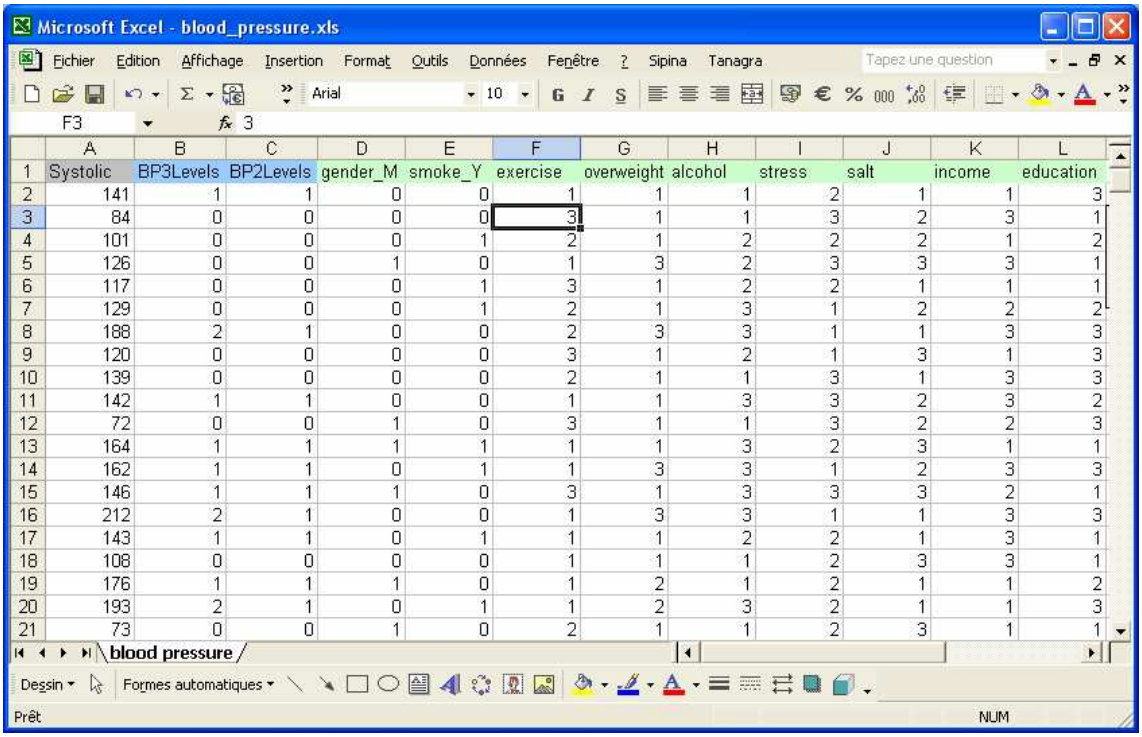
[1] http://www.math.yorku.ca/Who/Faculty/Ng/ssc2003/BPMain.htm

Stress	Stress level (1 : low ; 2 : medium ; 3 : high)
Salt	Salt intake level (1 : low ; 2 : medium ; 3 : high)
Income	Income level (1 : low ; 2 : medium ; 3 : high)
Education	Education level (1 : low ; 2 : medium ; 3 : high)

Observations

The original dataset contains patients which are treated and not treated for high blood pressure. When we use the whole dataset, only the treatment variable is significant for the prediction. It is not really interesting if the aim of the analysis is to detect the cause of the hypertension. So, we handle only the untreated patients in this tutorial (399 examples).

In the following screenshot, we display the first 20 examples of the dataset.



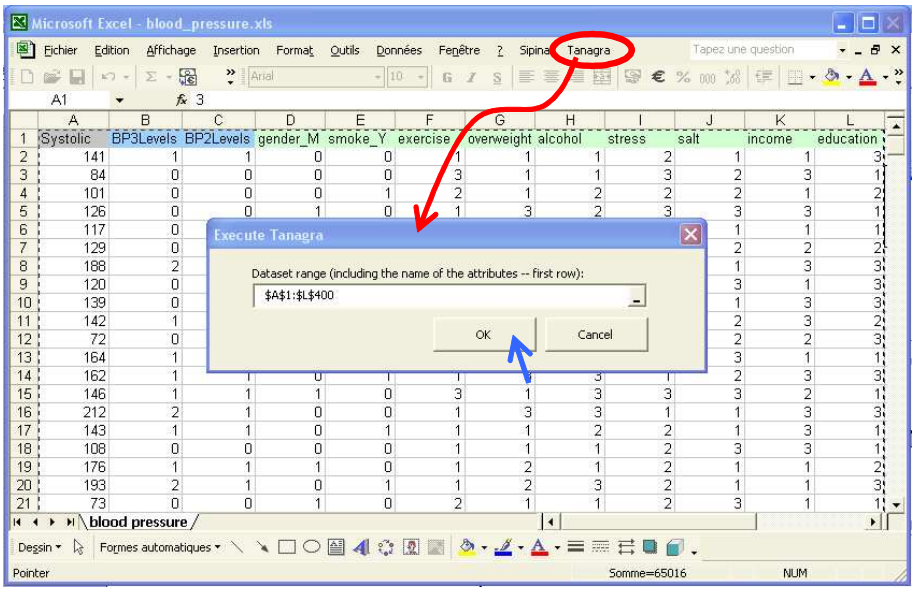
	A	B	C	D	E	F	G	H	I	J	K	L
1	Systolic	BP3Levels	BP2Levels	gender_M	smoke_Y	exercise	overweight	alcohol	stress	salt	income	education
2	141	1	1	0	0	1	1	1	2	1	1	3
3	84	0	0	0	0	3	1	1	3	2	3	1
4	101	0	0	0	1	2	1	2	2	2	1	2
5	126	0	0	1	0	1	3	2	3	3	3	1
6	117	0	0	0	1	3	1	2	2	1	1	1
7	129	0	0	0	1	2	1	3	1	2	2	2
8	188	2	1	0	0	2	3	3	1	1	3	3
9	120	0	0	0	0	3	1	2	1	3	1	3
10	139	0	0	0	0	2	1	1	3	1	3	3
11	142	1	1	0	0	1	1	3	3	2	3	2
12	72	0	0	1	0	3	1	1	3	2	2	3
13	164	1	1	1	1	1	1	3	2	3	1	1
14	162	1	1	0	1	1	3	3	1	2	3	3
15	146	1	1	1	0	3	1	3	3	3	2	1
16	212	2	1	0	0	1	3	3	1	1	3	3
17	143	1	1	0	1	1	1	2	2	1	3	1
18	108	0	0	0	0	1	1	1	2	3	3	1
19	176	1	1	1	0	1	2	1	2	1	1	2
20	193	2	1	0	1	1	2	3	2	1	1	3
21	73	0	0	1	0	2	1	1	2	3	1	1

Measures of Association for Ordinal Variables

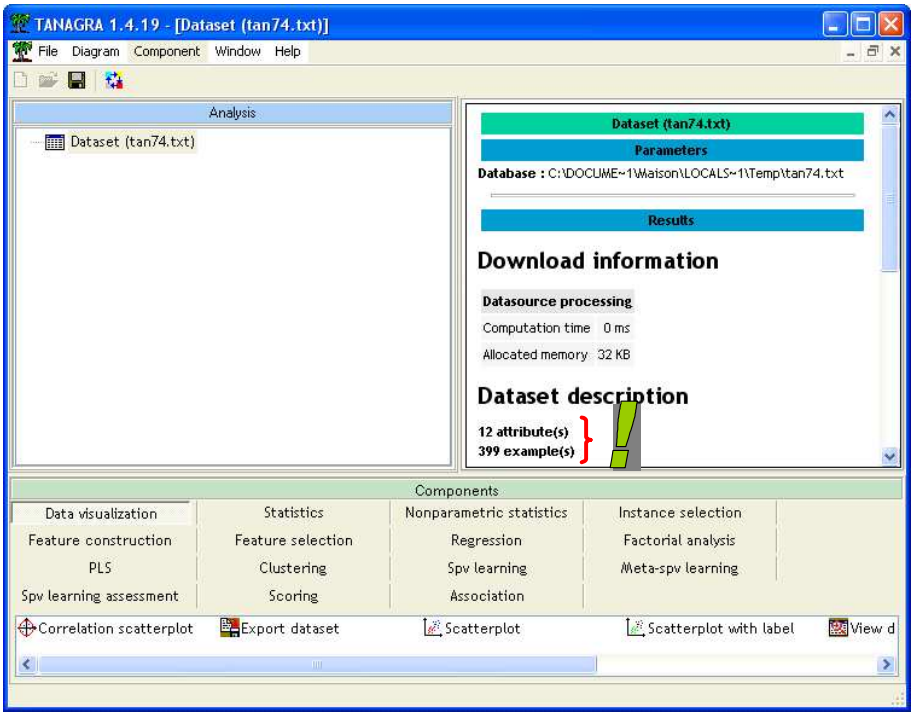
Create a new diagram

The simplest way in order to create a diagram is to load the dataset in the EXCEL spreadsheet. Then, we select the data range and we click on the menu TANAGRA/EXECUTE TANAGRA[2]. We check the range selection and we click on OK.

[2] The EXCEL add-in TANAGRA.XLA is available since the version 1.4.11. See the tutorial on the web site for the installation of this add-in in your spreadsheet.

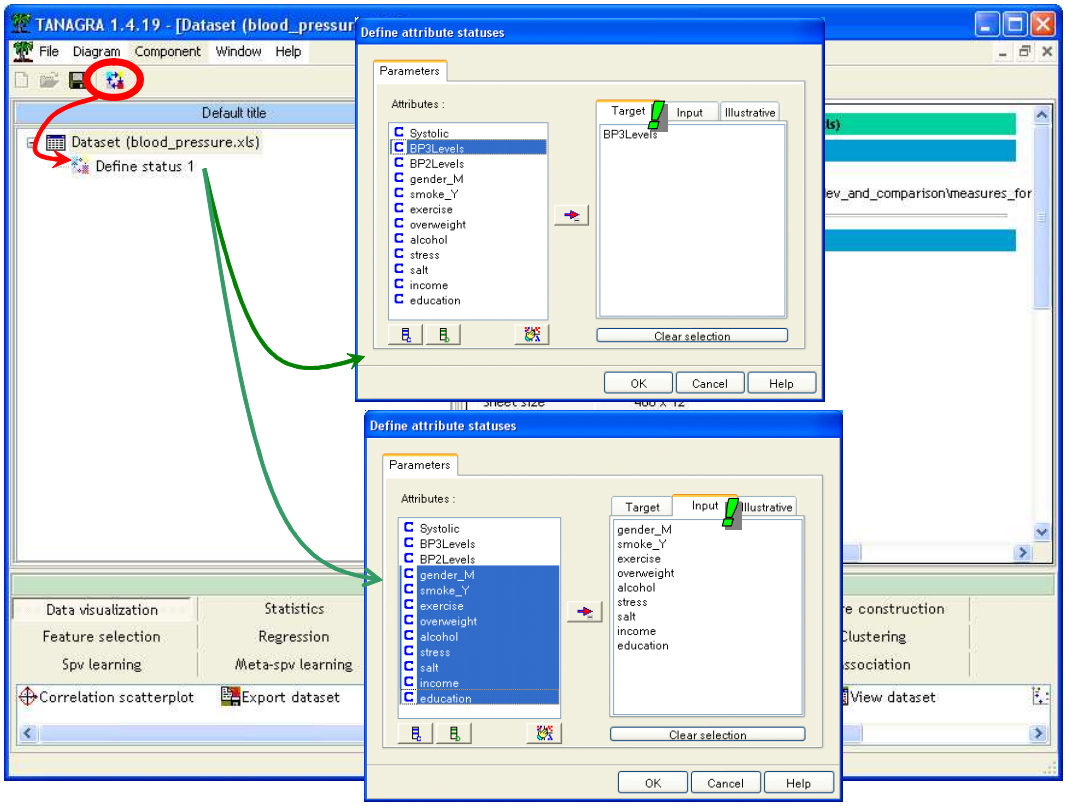


TANAGRA is automatically started, a new diagram is created and the data loaded. We check that the file contains really 399 observations and 12 variables.



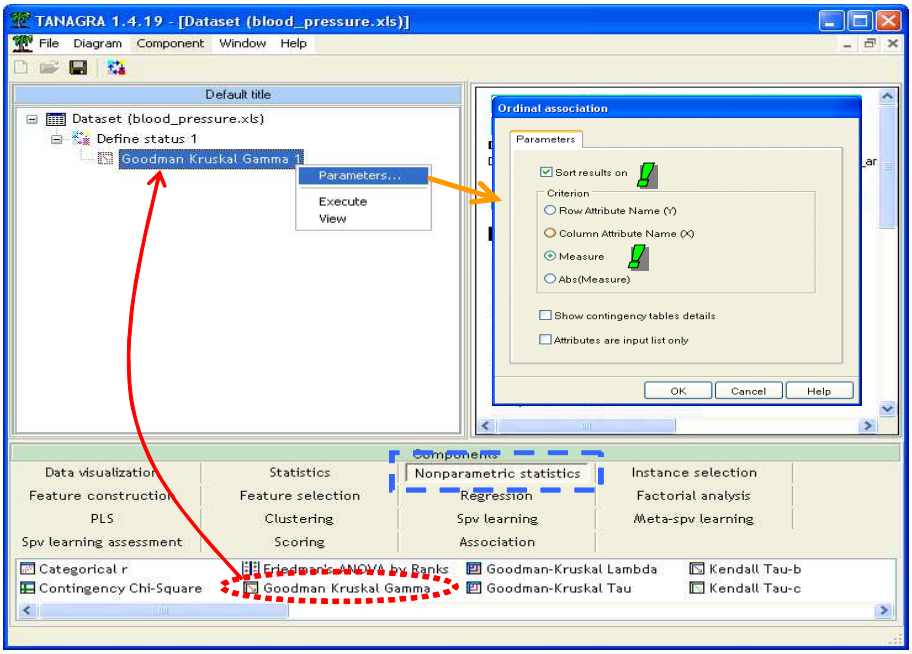
Explaining the BP3LEVES dependent variable

In the first time, we want to use the **Goodman and Kruskal's Gamma** measure. To do that, we must set BP3LEVELS as the TARGET variable, and the other one (GENDER_M to EDUCATION) as the INPUT variables. We add the DEFINE STATUS component in the diagram using the shortcut in the toolbar.

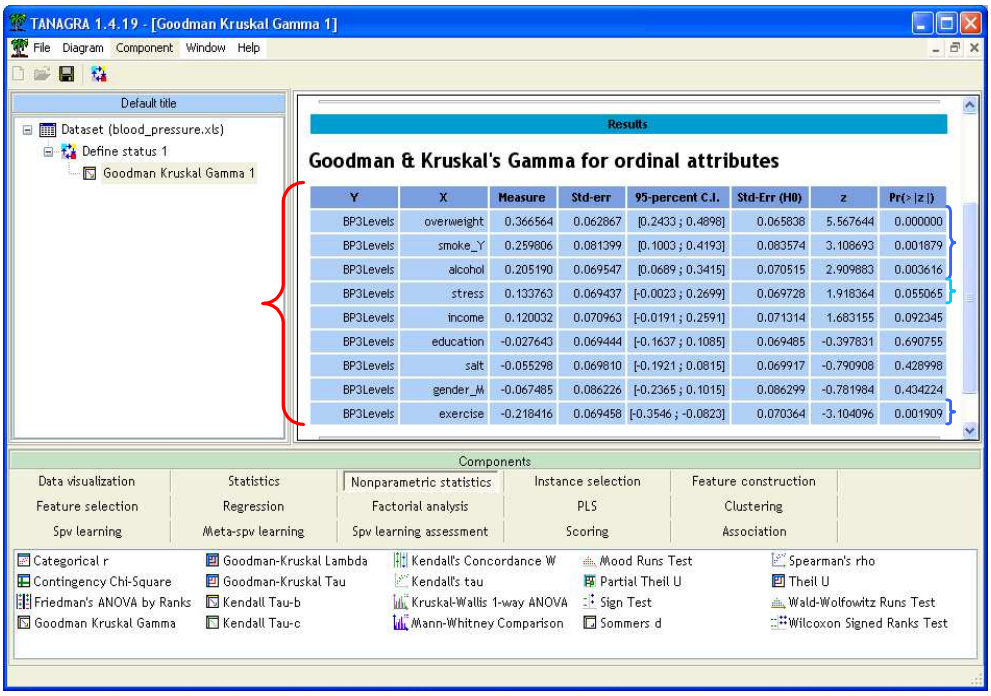


Note: Variable coding. We must use the code of the ordinal variables i.e. we set the values 1, 2, 3... and not the corresponding description (normal, high, very high).

We add into the diagram the Goodman Kruskal Gamma component (NONPARAMETRIC STATISTICS). We click on the Parameters menu in order to sort the results according the measure.



We obtain the results by clicking on the VIEW menu.

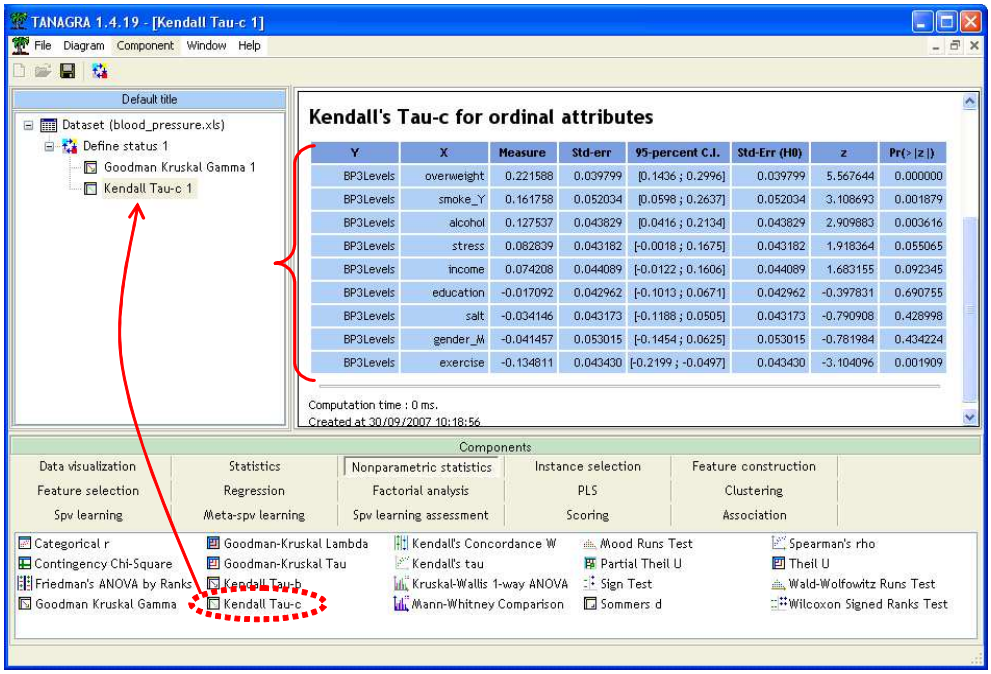


On each row of the table, we have the dependent and the independent variables, the measure, the standard error of the measure, the confidence interval for 95% level, the standard error under the independence assumption, the z statistic of the significance test, and the p-value of the test of independence. In our dataset, OVERWEIGHT, SMOKE_Y and ALCOHOL have a significant positive association with the dependent variable. EXERCISE has a negative association.

For the interpretation of the Gamma value, see the following reference - http://www2.chass.ncsu.edu/garson/PA765/assocordinal.htm.

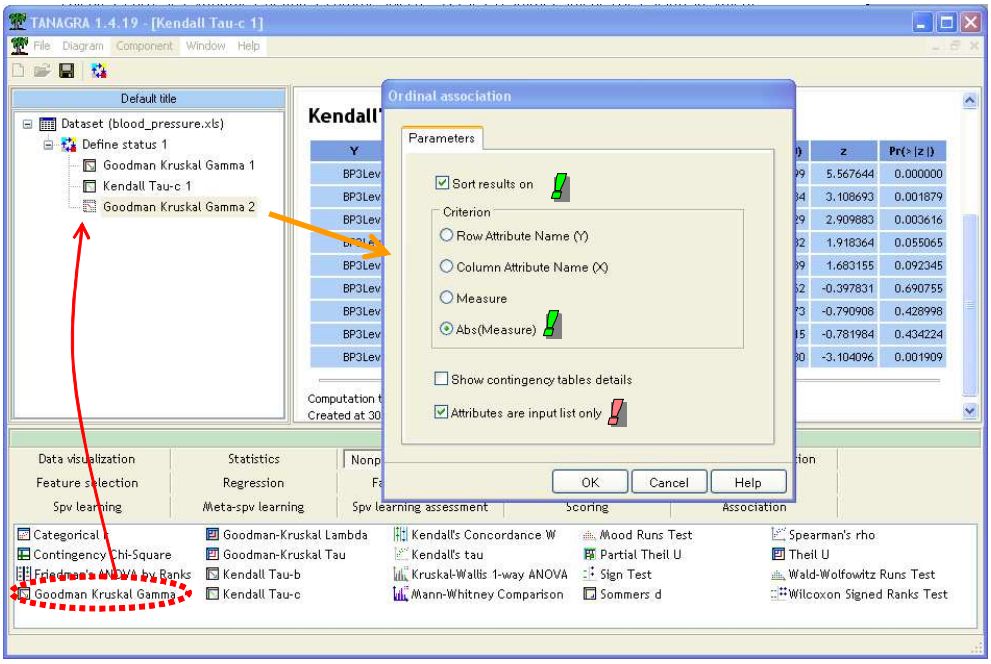
The other symmetrical measures

Kendall's Tau-b and Tau-c are also symmetrical measures. About the last one, we obtain the following results.



Inter-association between the input variables

Some variables have a high association with the hypertension. But perhaps, some of them are redundant and give the same kind of information. In order to detect these situations, we compute the association between the input variables. We use the same component i.e. Goodman Kruskal Gamma, but we slightly modify the parameters.



We obtain the following results.

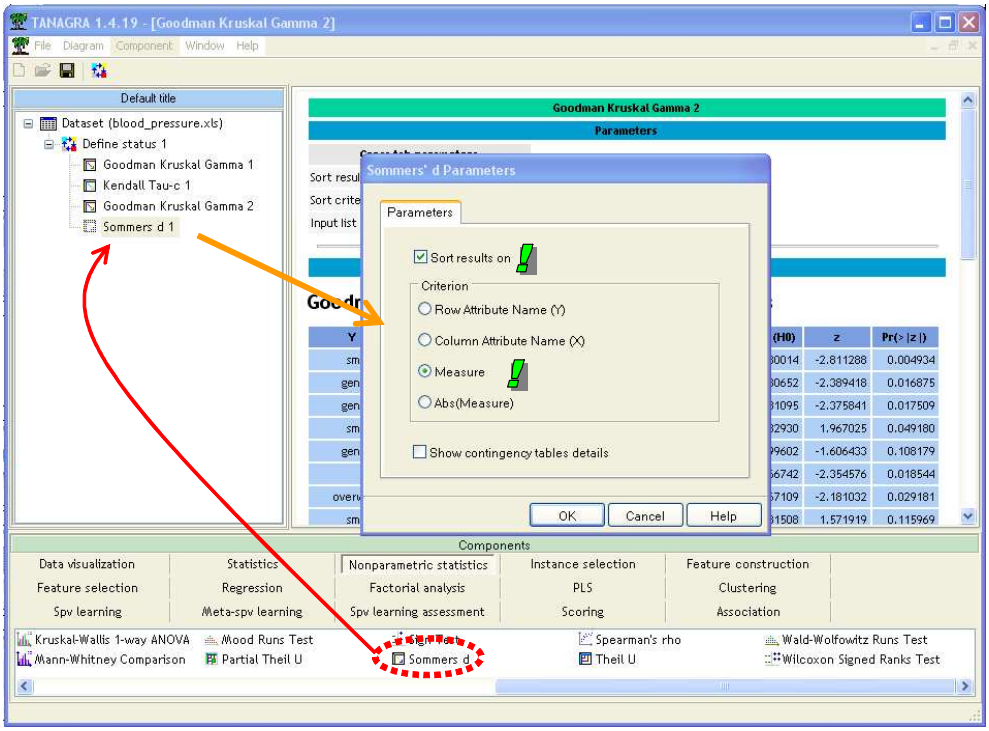
Goodman & Kruskal's Gamma for ordinal attributes

Y	X	Measure	Std-err	95-percent C.I.	Std-Err (H0)	z	Pr(>|z|)
smoke_Y	income	-0.224944	0.078479	[-0.3788 ; -0.0711]	0.080014	-2.811288	0.004934
gender_M	education	-0.192711	0.079493	[-0.3485 ; -0.0369]	0.080652	-2.389418	0.016875
gender_M	alcohol	-0.192669	0.079973	[-0.3494 ; -0.0359]	0.081095	-2.375841	0.017509
smoke_Y	overweight	0.163125	0.082	[0.0024 ; 0.3238]	0.08293	1.967025	0.04918
gender_M	smoke_Y	-0.160004	0.098273	[-0.3526 ; 0.0326]	0.099602	-1.606433	0.108179
salt	education	-0.157149	0.066271	[-0.2870 ; -0.0273]	0.066742	-2.354576	0.018544
overweight	alcohol	-0.146367	0.066696	[-0.2771 ; -0.0156]	0.067109	-2.181032	0.029181
smoke_Y	exercise	0.128124	0.080993	[-0.0306 ; 0.2869]	0.081508	1.571919	0.115969
stress	income	0.104521	0.065635	[-0.0241 ; 0.2332]	0.065841	1.587474	0.112405
exercise	salt	0.092326	0.067841	[-0.0406 ; 0.2253]	0.06802	1.357329	0.174677
gender_M	income	0.089839	0.081352	[-0.0696 ; 0.2493]	0.081617	1.100739	0.27101
alcohol	salt	-0.083667	0.067707	[-0.2164 ; 0.0490]	0.067859	-1.23296	0.217591
gender_M	exercise	-0.083594	0.081649	[-0.2436 ; 0.0764]	0.081857	-1.021213	0.307153
exercise	income	0.074749	0.065819	[-0.0543 ; 0.2038]	0.065926	1.133836	0.256863
smoke_Y	stress	0.071663	0.081441	[-0.0880 ; 0.2313]	0.081601	0.878209	0.37983
gender_M	stress	0.060967	0.082059	[-0.0999 ; 0.2218]	0.082165	0.742014	0.458079
alcohol	income	0.058377	0.068778	[-0.0764 ; 0.1932]	0.068841	0.848003	0.396436
smoke_Y	salt	-0.057968	0.081711	[-0.2181 ; 0.1022]	0.081808	-0.708581	0.478584
income	education	-0.054395	0.067248	[-0.1862 ; 0.0774]	0.067309	-0.808139	0.419011
stress	salt	-0.051508	0.067422	[-0.1837 ; 0.0806]	0.067474	-0.763384	0.445235
smoke_Y	alcohol	-0.05094	0.081703	[-0.2111 ; 0.1092]	0.081782	-0.622872	0.533369
smoke_Y	education	-0.044758	0.081714	[-0.2049 ; 0.1154]	0.081776	-0.547326	0.584155
exercise	education	-0.043204	0.067827	[-0.1761 ; 0.0897]	0.067864	-0.636622	0.524371
alcohol	education	-0.030246	0.066611	[-0.1608 ; 0.1003]	0.066627	-0.453965	0.649854
gender_M	salt	-0.02913	0.08211	[-0.1901 ; 0.1318]	0.082142	-0.354629	0.722868
overweight	salt	-0.028534	0.06827	[-0.1623 ; 0.1053]	0.068285	-0.417873	0.67604
overweight	income	-0.027911	0.067649	[-0.1605 ; 0.1047]	0.067666	-0.412481	0.679987
exercise	alcohol	-0.027764	0.067382	[-0.1598 ; 0.1043]	0.067401	-0.411916	0.680401
exercise	overweight	0.01409	0.068219	[-0.1196 ; 0.1478]	0.068222	0.206525	0.836381
overweight	stress	0.013756	0.069394	[-0.1223 ; 0.1498]	0.069398	0.198219	0.842873
gender_M	overweight	0.01308	0.08425	[-0.1520 ; 0.1782]	0.084257	0.155235	0.876636
overweight	education	0.010495	0.068828	[-0.1244 ; 0.1454]	0.068829	0.152474	0.878813
stress	education	0.009795	0.066509	[-0.1206 ; 0.1401]	0.06651	0.147271	0.882918
exercise	stress	0.007457	0.066921	[-0.1237 ; 0.1386]	0.066921	0.111426	0.911279
alcohol	stress	0.00543	0.067773	[-0.1274 ; 0.1383]	0.067773	0.080118	0.936143
salt	income	0.001247	0.066938	[-0.1299 ; 0.1324]	0.066938	0.018635	0.985132

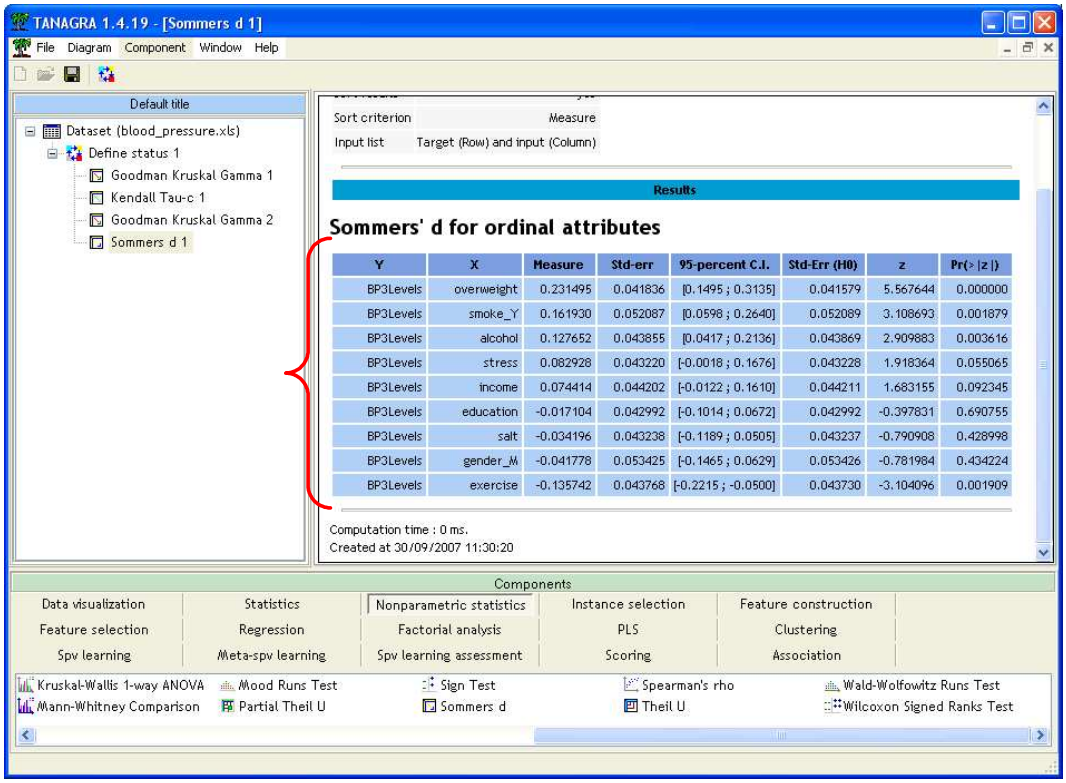
There are some significant associations between several INPUT variables. But, to really understand the signification of these associations and their influence on the hypertension, we need an expert domain.

Asymmetric association

We want to explain the hypertension with some independent variables, but at this time, we use symmetrical measures. If we transpose the contingency table, we obtain the same value. They cannot really detect directional association. In this section, we study another measure which is asymmetric: the Sommers' d. We add this component into the diagram.



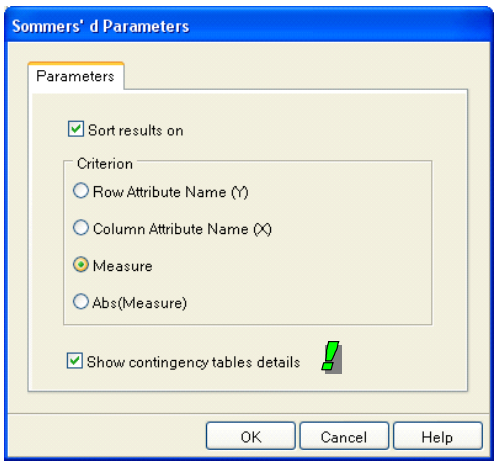
The value of the coefficient is different, but the significance test gives the same results.



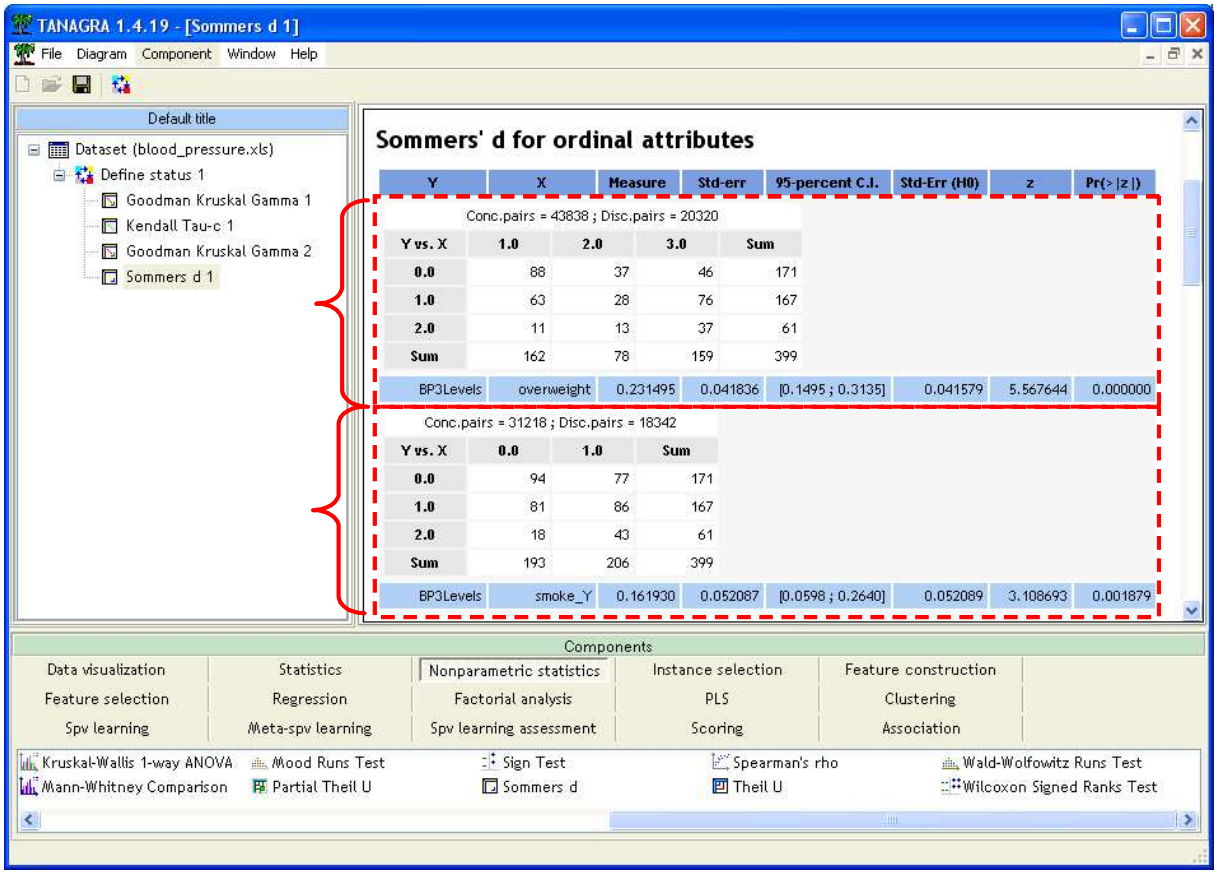
Details about the contingency table

In some situations, it may be useful to display the details of the contingency table e.g. when we want to characterize the kind of the association (monotonic, weak monotonic, etc. -- see http://www2.chass.ncsu.edu/garson/PA765/association.htm).

We click on the PARAMETERS menu of SOMMERS D 1 in the diagram, and we select the right option.



We display here the details of the contingency table for the 2 first associations (BP3LEVELS vs. OVERWIEGHT and BP3LEVELS vs. SMOKE_Y).



For instance, for BP3LEVELS vs. OVERWEIGHT, they are 43838 concordant pairs and 20320 discordant pairs.

Using the 2 levels dependent variable

When we discretize a continuous attribute for this kind of analysis, the transformation process may heavily influence the results. We try another transformation of the SYSTOLIC variable in 2 levels (BP2LEVELS, upper or lower than 140 mm hg).

Using the Goodman and Kruskal measure, we obtain another result about the STRESS variable.

Conc.pairs = 37260 ; Disc.pairs = 28468

Y vs. X	1.0	2.0	3.0	Sum
0.0	63	58	50	171
1.0	45	60	62	167
2.0	17	22	22	61
Sum	125	140	134	399

BP3Levels	stress	0.133763	0.069437	[-0.0023 ; 0.2699]	0.069728	1.918364	0.055065

Tab. 1 - BP3Levels vs. STRESS

Conc.pairs = 30660 ; Disc.pairs = 21592

Y vs. X	1.0	2.0	3.0	Sum
0.0	63	58	50	171
1.0	62	82	84	228
Sum	125	140	134	399

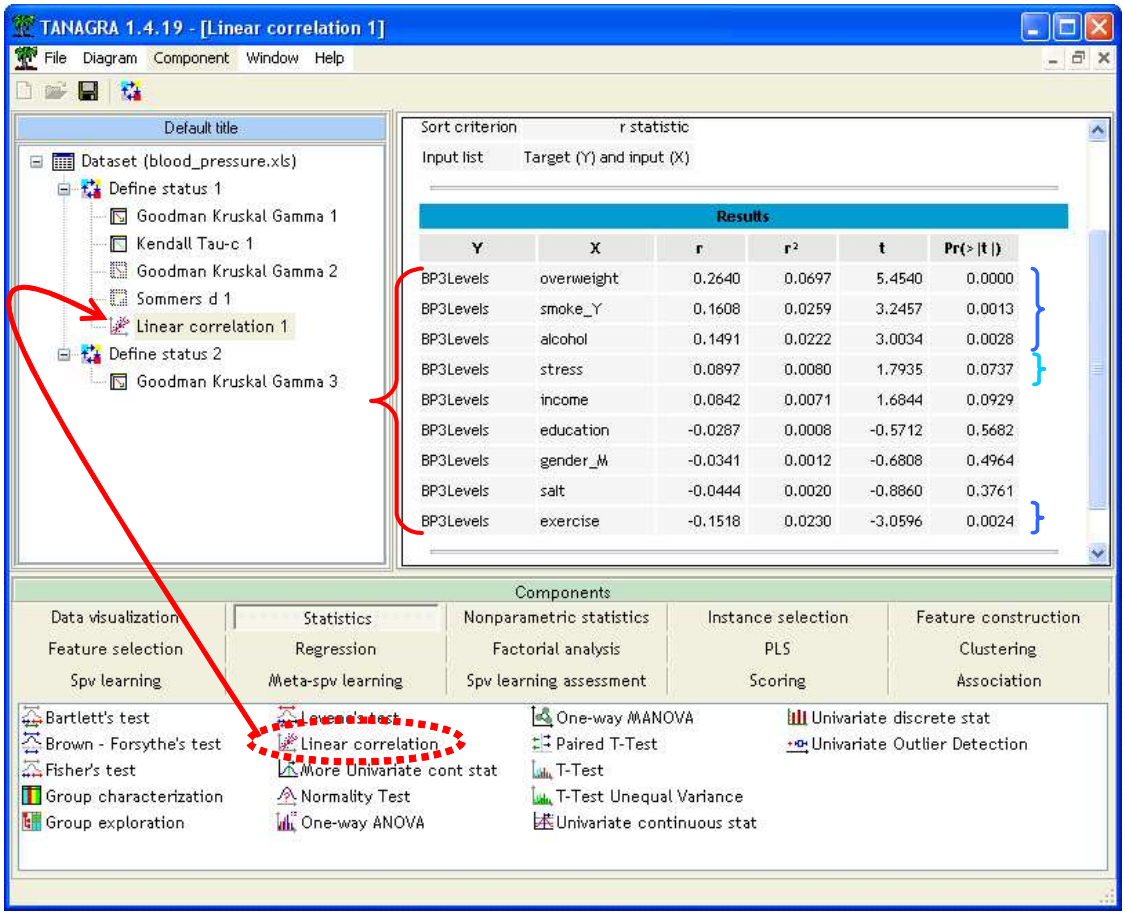
BP2Levels	stress	0.173544	0.080778	[0.0152 ; 0.3319]	0.081750	2.122848	0.033767

Tab. 2 - BP2Levels vs. STRESS

Comparing the ordinal measure with the correlation coefficient

The Pearson's correlation coefficient is not theoretically suitable for ordinal variables because we have not the right information about the scale of the various levels of the ordinal variable. We want to check this assertion on our dataset. We use the trivial values 1, 2, 3... for all the variables.

We add the LINEAR CORRELATION component into the diagram (STATISTICS tab).



Surprisingly (or not), we obtain similar results to the Gamma measure. Perhaps, the correlation coefficient is not so a very bad measure for ordinal variables if we use “reasonable” values (and not “exotic” values such as 1, 2000, 2001, etc.). Perhaps also, it is due to the fact that the majority of the variables in this analysis are discretized variables.

Conclusion

In this tutorial, we show how to use TANAGRA (1.4.19 and higher) for measuring the association between ordinal variables. The available measures are: Goodman and Kruskal Gamma, Kendall’s Tau-b and Tau-c, Sommers’ d.