# 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: <u>http://www2.chass.ncsu.edu/garson/PA765/association.htm</u>.

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

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<sup>1</sup>. 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)

<sup>&</sup>lt;sup>1</sup> <u>http://www.math.yorku.ca/Who/Faculty/Ng/ssc2003/BPMain.htm</u>

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.

	Aicrosoft Ex	ccel - blood	_pressure.	xls								
B)	Eichier Edi	tion <u>A</u> fficha	ge Insertio	n Forma <u>t</u>	<u>O</u> utils <u>D</u> or	nnées Fe <u>n</u> é	itre <u>?</u> Sip	ina Tanagr	a	Tapez une	question	• _ 8 ×
р		F	a »	Arial	• 1	0 - 6	<i>r</i> s ≡	三 三 扇		% 000 +20	se on .	. A . »
	E2	2 4	.⊃. ⊪⊡ •:.		100		1 <u>0</u> =-			/0 000 ,00		
-	r.j	• /·	* 3	D	F	Ê	0	Ц		E.	12	
1	Rustalia	D DOL ouolo	PD1 avala	acador M	E amaka V	F.	G	alaahal	otrago	J	incomo	aduaction 🔺
2	3ystolic 1/1	DFJLevels 1	DFZLEVEIS 1	gender_im	SINUKE_1	exercise 1	overweight 1	alconol 1	SHESS	San 1	income 1	education 3
2	141	0	0	0	0	3	1	1	2	2	3	1
1	101	0	0	0	1	2	1	2	2	2	1	2
5	126	0	0	1	'n	1	3	2	3	3	3	1
6	117	ñ	Ő	ò	1	3	1	2	2	Ĭ	1	1
7	129	Ő	ŏ	Ő	1	2	i i	3	1	2	2	2
8	188	2	1	Ō	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 -
14 4	> > > >> >> >> >> >> >>>>>>>>>>>>>>>>>	od pressure	e/					1			1	<u>)</u> ]
Deg	isin 🕶 🔓 🛛 Fa	ormes automat	tiques 🕶 🔨	× 🗆 O			🔊 - 🔏 - 🛔	▲ - 〓 =		0 -		
Prêt											NUM	

### 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<sup>2</sup>. We check the range selection and we click on OK.

 $<sup>^2</sup>$  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.

3	Eichier Ed	lition	Affichage	Insertion	Forma <u>t</u>	Outils	Données	Fenêtr	e <u>?</u> Sipi	ina Tanagi	a	Tape	z une q	uestion	
0	- 1	0 - 0	Σ - 🚱	* A	nal		- 10 -	G 2	s 🗐			% 000	:28	(F 🛄 •	🕭 - 🛕
	A1	•	fx :	3											
	A	E	3	С	D	E		=	G	Н	1	J		K	L
1	Systolic	BP3L	evels BF	P2Levels	gender_M	smoke_	Y exerc	cise o	verweight	alcohol	stress	salt	ir	ncome	education
2	141		1	1	0		0	1	1	1		2	1	1	3
3	84		0	0	0		0	3	1	্		3	2	3	1
1	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
5	117		0	Execute	e Tanagra							X	1	1	1
1	129		0	CONCEPTIONS									2	2	2
	188		2	D	ataset range	(including	the name	of the a	ttributes f	irst row):			1	3	2
)	120		0	Ē	#0#1.#1#4		the mane	or and a		a per con yr	1		3	1	3
0	139		0		рмр1.рсртс	0					_		1	3	6
1	142		1							4	5 1		2	3	2
2	72		0						OK	Cano	el		2	2	3
3	164		1										3	1	
4	162		1	1	U		1	1	Ŭ,	3		1	2	3	1
5	146		1	1	1		0	3	1	3		3	3	2	
3	212		2	1	0		0	1	3	3		1	1	3	
7	143		া	1	0		1	1	2	2	8	2	1	3	
8	108		0	0	0		0	1	1	1		2	3	3	1
9	176		1	1	1		0	1	2	1		2	1	1	2
D	193		2	1	0		1	1	2	3		2	1	1	3
1	73		0	0	1		0	2	1	1		2	3	1	1
.4	→ → \ blc	ood pre	essure /							1					•

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.

TANAGRA 1.4.19 - [Dat	aset (tan74.txt)]			
Tile Diagram Component	Window Help			_ 8 ×
D 📽 🖪 🔯				
	Analysis		Dataset dan 74 tub	
📃 🥅 Dataset (tan74.txt)			Parameters	
		Database : C:\DOC	CUME~1Waison\LOCALS~1\Temp\tan?	74.txt
			10 10	
			Results	
		Download	information	
		Datasource proc		
		Computation time		
		Allocated memory	32 KB	
		Dataset de	scription	
		12 attribute(s)	. [/	
		399 example(s)		~
		Components		
Data visualization	Statistics	Nonparametric statistics	Instance selection	
Feature construction	Feature selection	Regression	Factorial analysis	
PLS	Clustering	Spv learning	Meta-spv learning	
Spv learning assessment	Scoring	Association		
Correlation scatterplot	Export dataset	🚀 Scatterplot	🛃 Scatterplot with label	🔣 View d
2				>
			2	

### 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.

File Diagram Component W	/indow Help								-		
-   📽 🖪   🎎											
Default title						1.0000					
🖃 🎹 Dataset (blood_pressur	e.xls)				Res	ults					
🖻 🚰 Define status 1	al Gamma 1	Goodman & Kruskal's Gamma for ordinal attributes									
		Y	x	Measure	Std-err	95-percent C.I.	Std-Err (H0)	z	Pr(> z )		
		<b>BP3Levels</b>	overweight	0.366564	0.062867	[0.2433; 0.4898]	0.065838	5.567644	0.000000		
		<b>BP3Levels</b>	smoke_Y	0.259806	0.081399	[0.1003;0.4193]	0.083574	3.108693	0.001879		
		<b>BP3Levels</b>	alcohol	0.205190	0.069547	[0.0689 ; 0.3415]	0.070515	2.909883	0.003616		
		BP3Levels	stress	0.133763	0.069437	[-0.0023; 0.2699]	0.069728	1.918364	0.055065		
		BP3Levels	income	0.120032	0.070963	[-0.0191 ; 0.2591]	0.071314	1.683155	0.092345		
		BP3Levels	education	-0.027643	0.069444	[-0.1637; 0.1085]	0.069485	-0.397831	0.690755		
		BP3Levels	salt	-0.055298	0.069810	[-0.1921; 0.0815]	0.069917	-0.790908	0.428998		
		BP3Levels	gender_M	-0.067485	0.086226	[-0.2365; 0.1015]	0.086299	-0.781984	0.434224		
		BP3Levels	exercise	-0.218416	0.069458	[-0.3546 ; -0.0823]	0.070364	-3, 104096	0.001909		
			Comp	onents							
Data visualization	Statistics	Nonpara	metric statistics	Instar	nce selecti	on Featur	e construction				
Feature selection	Regression	Fact	orial analysis		PLS	0	lustering				
Spv learning	Meta-spv learning	Spv lear	ning assessment		Scoring	A	sociation				
Categorical r	🗐 Goodman-Kruskal	Lambda	🗄 Kendall's Conc	ordance W		d Runs Test	🖄 Spear	man's rho			
Contingency Chi-Square	🗾 Goodman-Kruskal	Tau	🖄 Kendall's tau		🛱 Parti	al Theil U	🗾 Theil	U			
Friedman's ANOVA by Ranks	🖾 Kendall Tau-b		🛴 Kruskal-Wallis	1-way ANOVA	Sign	Test	🚢 Wald-	Wolfowitz	Runs Test		
] Goodman Kruskal Gamma	🔟 Kendall Tau-c	L.	🗽 Mann-Whitney	Comparison	🗖 Somr	ners d	<b>∷</b> #Wilco	xon Signed	i Ranks Test		

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 - <u>http://www2.chass.ncsu.edu/garson/PA765/assocordinal.htm</u>.

#### The other symmetrical measures

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

🕈 TANAGRA 1.4.19 - [Ken	dall Tau-c 1]												
File Diagram Component	Window Help								-	6			
) 📽 🖪  🍇													
Default title										ĺ			
🖃 🥅 Dataset (blood_press)	ure.xls)	Kendall's Tau-c for ordinal attributes											
🖃 🚰 Define status 1		Y	X	Measure	Std-err	95-percent C.I.	Std-Err (H0)	z	Pr(> z )				
Goodman Krus	skal Gamma 1	BP3Levels	overweight	0.221588	0.039799	[0.1436; 0.2996]	0.039799	5.567644	0.000000				
Kendall Tau-c		<b>BP3Levels</b>	smoke_Y	0.161758	0.052034	[0.0598; 0.2637]	0.052034	3,108693	0.001879				
Ť		BP3Levels	alcohol	0.127537	0.043829	[0.0416; 0.2134]	0.043829	2.909883	0.003616				
		<b>BP3Levels</b>	stress	0.082839	0.043182	[-0.0018; 0.1675]	0.043182	1.918364	0.055065				
	$\prec$	<b>BP3Levels</b>	income	0.074208	0.044089	[-0.0122; 0.1606]	0.044089	1.683155	0.092345				
		<b>BP3Levels</b>	education	-0.017092	0.042962	[-0.1013; 0.0671]	0.042962	-0.397831	0.690755				
		<b>BP3Levels</b>	salt	-0.034146	0.043173	[-0.1188; 0.0505]	0.043173	-0.790908	0.428998				
		<b>BP3Levels</b>	gender_M	-0.041457	0.053015	[-0.1454; 0.0625]	0.053015	-0.781984	0.434224				
		<b>BP3Levels</b>	exercise	-0.134811	0.043430	[-0.2199; -0.0497]	0.043430	-3.104096	0.001909				
		715 452 105	12										
	Con	nputation time	: 0 ms. /2007 10:18:56										
		atou at 30/07	72007 10.10.50							-			
Data visualization	Statistics	Nonpara	metric statistics	Insta	nce selecti	ion Featur	e construction	n l'					
Feature selection	Regression	Facto	nrial analysis		PLS	0	lustering						
Spy learning Meta-spy learning Spy		Spv learr	ning assessment		Scoring	A	ssociation						
Categorical r	🔲 Goodman-Kruskal	Lambda 🖡	🖞 Kendall's Conc	ordance W	iii, Moo	d Runs Test	🖄 Spe	arman's rho		_			
Contingency Chi-Square	il Tau 🖉 Kendall's tau			🛱 Parti	ial Theil U	🗾 The	il U						
Friedman's ANOVA by Rank	ks 🛛 Kendall Tau-b	. la	Kruskal-Wallis	1-way ANOV	A 📑 Sign	Test	🚉 Wald	J-Wolfowitz	Runs Test				
🔄 Goodman Kruskal Gamma	🔪 🗋 Kendall Tau-c		🗽 /Mann-Whitney	Comparisor	n 🖬 Somi	mers d	It Wilc	oxon Signed	1 Ranks Test				
										_			

## 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	Х	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.

FIRMAGKA 1.4.19 - [Goodman Krus File Diagram Component Window H	skat Gamma <b>2</b> ] Sp							
Default title			Goodman Kruskal Ga	nma 2				
<ul> <li>□ Imataset (blood_pressure.xls)</li> <li>□ Imata</li></ul>	1 Sort resul	Sommers' d Parameter	Parameters 5					
- 🖸 Goodman Kruskal Gamma - 🖾 Sommers d 1	2 Sort crite	Parameters Sort results on Criterion	8					
	Y sm gen gen	Row Attribute     O Column Attrib     O Measure     Abs(Measure	Name (Y) ute Name (X) )		(H0) 30014 30652 31095	z -2.811288 -2.389418 -2.375841	Pr(> z ) 0.004934 0.016875 0.017509	
	sm gen overs-	Show continge	ency tables details		32930 99602 36742 37109	1.967025 -1.606433 -2.354576 -2.181032	0.049180 0.108179 0.018544 0.029181	
	sm		OK Cancel	Help	31508	1.571919	0.115969	
Data visualization Sta Feature selection Reg Spv learning Meta-sp	tistics Nor ression w.learning Spw	Compone nparametric statistics Factorial analysis / learning assessment	ents Instance selection PLS Scoring	Feature con Cluste Associa	structior ring ition	1		
Kruskal-Wallis 1-way ANOVA 🛛 🗮 Moo Mann-Whitney Comparison 🛛 🕅 Part	d Runs Test ial Theil U	Sommers d	🖄 Spearman's r 🖭 Theil U	o	iii, Wald ∷≌Wilc	1-Wolfowitz .oxon Signer	Runs Test 1 Ranks Test	

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

File Diagram Component Window Help											
📽 🖬 🙀											
Default title IIII Dataset (blood_pressure.xks) IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	Sort criterio Input list	n Target (Row) and	, Measure input (Column)								
- 🔽 Goodman Kruskal Gamma 1 - 🗖 Kendall Tau-c 1 - 🖸 Goodman Kruskal Gamma 2 - 📮 Sommers d 1	Somme	Results Sommers' d for ordinal attributes									
	Y	х	Measure	Std-err	95-percent C.I.	Std-Err (H0)	z	Pr(> z )			
	BP3Lev	vels overweight	0.231495	0.041836	[0.1495; 0.3135]	0.041579	5.567644	0.000000			
	BP3Lev	/els smoke_Y	0.161930	0.052087	[0.0598; 0.2640]	0.052089	3.108693	0.001879			
	BP3Lev	vels alcoho	0.127652	0.043855	[0.0417; 0.2136]	0.043869	2.909883	0.003616			
	BP3Lev	vels stress	0.082928	0.043220	[-0.0018; 0.1676]	0.043228	1.918364	0.055065			
	BP3Lev	/els income	0.074414	0.044202	[-0.0122; 0.1610]	0.044211	1.683155	0.092345			
	BP3Lev	education	-0.017104	0.042992	[-0.1014; 0.0672]	0.042992	-0.397831	0.690755			
	BP3Lev	rels salt	-0.034196	0.043238	[-0.1189; 0.0505]	0.043237	-0.790908	0.428998			
	BP3Lev	/els gender_M	-0.041778	0.053425	[-0.1465; 0.0629]	0.053426	-0.781984	0.434224			
	BP3Les	rels exercise	-0.135742	0.043768	[-0.2215;-0.0500]	0.043730	-3,104096	0.001909			
	Computation Created at 30	time : 0 ms. )/09/2007 11:30:20									
		Con	nponents								
Data visualization Statistic	s Nont	varametric statistic	s Insta	ince selecti	ion Featur	e construction					
Feature selection Regression	in F	actorial analysis		PLS	(	Clustering					
Spv learning Meta-spv lea	rning Spv I	earning assessmen	t	Scoring	A	ssociation					
🥻 Kruskal-Wallis 1-way ANOVA 🛛 📠 Mood Ru Mann-Whitney Comparison 🛛 🛱 Partial TI	ns Test neil U	: Sign Test		🔄 Spea	arman's rho LU	<u>iiii,</u> Wald ∷ <mark>∺</mark> Wilco	-Wolfowitz oxon Signer	Runs Test d Ranks Test			
							userrende <del>s (</del> 1744)				

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 <a href="http://www2.chass.ncsu.edu/garson/PA765/association.htm">http://www2.chass.ncsu.edu/garson/PA765/association.htm</a>).

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

Sommers' d Parameters
Parameters
Sort results on
Criterion
O Row Attribute Name (Y)
◯ Column Attribute Name (X)
⊘ Measure
O Abs(Measure)
Show contingency tables details
OK Cancel Help

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

mail TANAGRA 1.4.19 - [Sommers d 1]									
💇 File Diagram Component Window Help								-	a ×
Default title           Default title           Image: Dataset (blood_pressure.xls)	Sommers'	d for ordi	nal attr	ibutes					^
🖻 🞇 Define status 1	Y	Х	Measure	Std-err	95-percent C.I.	Std-Err (H0)	z	Pr(> z )	
🖸 Goodman Kruskal Gamma 1	Con	nc.pairs = 43838 ; I	Disc.pairs =	20320					
Goodman Kruskal Gamma 2	Y vs. X	1.0 2.0	3.	) Su	m			i	
Sommers d 1	0.0	88	37	46	171				
$\sim$	1.0	63	28	76	167				_
	2.0	11	13	37	61			- i	
	Sum	162	78	159	399				
	BP3Levels	overweight	0.231495	0.041836	[0.1495; 0.3135]	0.041579	5.567644	0.000000	
	Conc.pair:	s = 31218 ; Disc.pa	airs = 18342						
	Y vs. X	0.0 1.0	Su	n				i	
	0.0	94	77	171					
$\prec$	1.0	81	86	167					
	2.0	18	43	61				- i	
	Sum	193	206	399					
L U	BP3Levels	smoke_Y	0.161930	0.052087	[0.0598; 0.2640]	0.052089	3.108693	0.001879	~
		Comp	onents						
Data visualization Statistics	Nonpara	metric statistics	Inst	ince selecti	on Featur	e construction	e l		
Feature selection Regression	Facto	orial analysis		PLS	0	lustering	0		
Spv learning Meta-spv learning	Spv learr	ning assessment		Scoring	A	ssociation	L.		
🗽 Kruskal-Wallis 1-way ANOVA 🛛 🛲 Mood Runs Te	st 🔤	Sign Test		🖄 Spea	rman's rho	🖮 Wald	-Wolfowitz F	Runs Test	
👫 Mann-Whitney Comparison 🛛 🛱 Partial Theil U	[	🛛 Sommers d		🗹 Thei	ιυ	⊑ <sup>‡‡</sup> Wilc	oxon Signed	Ranks Test	
<									>

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							
Yvs.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			
BP3Lev	els	stress 0.1	133763 0.0	69437 [-0.0			

Tab. 1 - BP3Levels vs. STRESS

	Conc.pairs	= 30660 ;	; Disc.,	pairs = 21						
Yvs.X	1.0	2.	0	3.0		Sum				
0.0	6	3	58		50	171				
1.0	63	2	82		84	228				
Sum	12	5	140	1	34	399				
BP2Lev	/els	stress	0.1	73544	0.0807	78 [0.0	ļ	152;0.3319]	152;0.3319] 0.081750	152;0.3319] 0.081750 2.122848

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.