# Subject

Measures of association for nominal variables.

To measure the association between two continuous variables, we generally use the correlation coefficient. Its drawbacks and its qualities are well known.

When we want to characterize the association for nominal variables, the correlation coefficient is not suitable. We must use other indicators. The most widespread is certainly the chi-square test, it enables to evaluate the absence of relation. We see in this tutorial that other measures are available. We show how to use them with TANAGRA.

Descriptions of the measures depicted in this tutorial are available on the following website:

- http://www.georgetown.edu/faculty/ballc/webtools/web\_chi\_tut.html
- http://v8doc.sas.com/sashtml/stat/chap28/sect20.htm
- http://www2.chass.ncsu.edu/garson/PA765/assocnominal.htm

#### Dataset

In the FUEL\_CONSUMPTION.XLS dataset, 205 cars are described by their CONSUMPTION, FUEL\_TYPE and ASPIRATION.

We show here the 20 first examples.

🖾 N	licrosoft Exce	el - fuel_cons	umption.xls				
	<u>Fi</u> chier <u>E</u> dition	Affichage Inse	rtion Forma <u>t</u> <u>C</u>	utils <u>D</u> onnées	Fe <u>n</u> être <u>?</u> Ta	inagra Sipina	_ 8 ×
	🚘 🗖 🖾	<u>}</u> #8∕ ¥ ⊑	a 🙉 🛷 🗠	1 × Ci + 🙆	Σ. f. 🗟 4	<u>, , , , , , , , , , , , , , , , , , , </u>	100% - » 💩 - »
10				1990	~ /~ 45 2		·····
	~		C	D	F	E	с н <del>–</del>
1	Consumption	fuel-type	aspiration		L	1	<u> </u>
2	high	nas	turbo				
3	high	gas	turbo				
4	high	gas	std				
5	high	gas	std				
6	high	gas	std				
7	high	gas	std				
8	high	gas	std				
9	high	gas	turbo				
10	high	gas	turbo				
11	high	gas	std				
12	high	gas	turbo				
13	high	gas	std				
14	high	gas	std				
15	high	gas	std				
16	high	gas	std				
17	high	gas	std				
18	high	gas	std				
19	high	gas	turbo				
20	high	gas	turbo				
21	high	gas	std				· · · · · · · · · · · · · · · · · · ·
	▶ ► \datas	et/					<b>&gt;</b>
Des	sin + 🗟 🌀	Formes automa	tiques 🗸 🔪 🍾		4 🗕 🔌 -	· <u>⊿</u> - <u>A</u> - ≡	∎ ☴ ◘ 🗊 -
Prêt	:						NUM

The goal is to measure the strength of the relation between ACCELERATION and the other variables.

## Chi Square test of independence

#### 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>1</sup>.

N 12	licrosoft Exce	l - fuel_cons	umption.xls								
	<u>Fi</u> chier <u>E</u> dition	Affichage Inse	rtion Forma <u>t</u>	<u>O</u> utik	s <u>D</u> onnées	Fe <u>n</u> être	e <u>2</u>	Tan	agra Sipina		_ 8 ×
	🖻 🖪 🍯	🐧 🚏 🐰 🗉	b 🛍 💅 🛛	<u>ب</u> (د)	CH 🗸 🎑	Σf	- <b>G</b> e		Execute Tanag	ra 🧲	🔫 🕭 - 🤻
-	C206 👻	= st	d						×		
	A	В	С		D		E		F	G	H
1	Consumption	fuel-type	aspiration								<b>_</b>
2	high	gas	turbo								
3	high	gas	turbo								
4	high	gas	std								
5	high	gas	std								
6	high	gas	std								
7	high	gas	std								
8	high	gas	std								
9	high	gas	turbo								
10	high	gas	turbo								
11	high	gas	std								
12	high	gas	turbo								
13	high	gas	std								
14	high	gas	std								
15	high	gas	std								
16	high	gas	std								
17	high	gas	std								
18	high	gas	std								
19	high	gas	turbo								
20	high	gas	turbo								
21	high	gas	std			L					<b>–</b>
	▶ ► \datas	et /					•				
Des	sin + 🗟 🕝	Formes automa	tiques 👻 🔨	×□		4 🙎	8	• •	<u>⊿</u> - <u>A</u> - ≣	= ☴ ☵ ᡕ	. 🗇 .
Prêt	:									NUM	

We check the range selection and we click on OK.



TANAGRA is automatically executed. We check that 3 variables and 205 examples are processed.

 $<sup>^{1}</sup>$  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.



#### Crosstabulation table and chi square statistic

We add the DEFINE STATUS component by using the short cut into the toolbar. First, we want to analyze the relation between CONSUMPTION (TARGET) and FULE-TYPE (INPUT).



We add the component CONTINGENCY CHI-SQUARE (NONPARAMETRIC tab) into the diagram and we click on the VIEW menu.

	RA 1.4.18 - [Co	ontingency Chi-	Square 1]								
		i milion noip									0 14
	Analysis	ĺ	Results								
🖃 🏢 Dat	taset (tan23C.tx Define status 1	t)	Row (Y)	Column (X)	Statistical	Statistical indicator			ross-tab		
L.	Contingency Chi-Square 1				Stat	Value		gas	diesel	Sum	
	1				Tschuprow's t	0.164399	high	41	0	41	
					Cramer's v	0.164399	low	144	20	164	
					Phi <sup>z</sup>	0.027027	Sum	185	20	205	
			Consumpti	on fuel-type	Chi² (p-value)	5.54 (0.0186)					
					Lambda	0.000000					
					Tau (p-value)	0.0270 (0.0189)					
					U(R/C) (p-value)	0.0461 (0.0021)					
			Computatio	Computation time : 0 ms.							
<u>р.                                    </u>			Created at	267057200205	*: /n: /n						
Data vi	walization	Statisti	20	Nonnaramet	ric statistics	Instance sel	ection	Fea	ture constri	uction	
Featur	selection	Regress	ion	Factorial	analysis	PLS			Clustering		
Sov	earning	Meta-spy le	arning	Spy learning	assessment	Scoring	2		Associatio	n l	
Categor	ical r	El Goodma	n-Kruskal Lan	bda istik	endall's tau	.in. 1	- Mood P	uns Test	t.e	<sup>7</sup> Spearmach	rbo
Contine	ency Chi-Square	Goodma	n-Kruskal Tau	idi Ki	ruskal-Wallis 1-wa	av ANOVA 🛛 🖾 F	Partial 7	Theil U		, spearman. I Theil U	
Friedma	in's ANOVA by R	anks 📲 Kendall	s Concordanc	e W 🚻 M	ann-Whitney Co	mparison 🏥 S	Sign Test 🗮 Wald-Wolfor				witz
											_
<							_				>

**Crosstabulation table.** CROSS-TAB is a contingency table. It allows examining the frequencies of observations that belong to each combination of categories of the variables. We note that there are 205 cars in the dataset, 20 of them use diesel fuel-type. Among these "FUEL-TYPE = diesel" cars, all have "CONSUMPTION = low".

**Chi-Square statistic.** The Pearson CHI-SQUARE (CHI-2) is the most common test for significance of the relationship for nominal variables. This measure is based on the fact that we can compute the *expected* frequencies in the contingency table (i.e., frequencies that we would *expect* if there was no relationship between the variables). Chi-square measures the deviation of the observed frequencies to the expected frequencies. It is 5.54 in our dataset.

The *p*-value of the test allows us to determine if you must accept or reject the null hypothesis. We can compare this value with a user-predefined significance level (p-level) of the test (e.g. 5% level in the most cases). In our dataset, the p-value is 0.0189. We reject the null hypothesis (reject "absence of relation") for a 5%-level.

🖬 Stats : CONSUMPT(2) x FUEL_TYP(2) (fuel_consumption.sta)				📷 Table Synthèse	à 2 Entrées : E	ffectifs Obs.	- 🗆 ×
STAT. ELEMENT.	Chi²	dl	p	STAT. ELEMENT.	Effect. ce	llules marc	µuées >10
Chi² de Pearson	5.540541	df=1	p=.01858		FUEL_TYP	FUEL_TYP	Totaux
Chi² du MV	9.452478	df=1	p=.00211	CONSUMPT	gas	diesel	Lignes
Phi des tables 2 x 2	.1643990		_	high : high	41	0	41
Corrél. tétrachorique	.5125254			low : low	144	20	164
Coeff. de contingence	.1622214			Tot. Colonnes	185	20 _	205

For comparison, here are the results of the STATISTICA software.

TANAGRA computes some additional statistical indicators derived from CHI-2, e.g. PHI-2, Tschuprow's t, Cramer's v.

**Chi-square contributions and calculations details.** It is possible to obtain the detail of calculations in the contingency table. We can for instance visualize the expected frequencies table:

we activate contextual menu PARAMETERS, and in ANALYSIS tab we click on the EXPECTED VALUES option.

TANAGRA 1.4.18 - [Da	ntaset (tan23C.)	txt)]	
The Diagram Component	t window help		^
		II	
Analysis		Cross-tabulation parameters	<u>^</u>
🖃 🏢 Dataset (tan23C.tx)	:)		
🖃 🎦 Define status 1	Chi Carros A	Database : C:\DOCUME~1 Parameters Analysis	
		Parameters       Additional information         Execute       None         View       Id         Datasource processing       Row percent         Computation time       0 ms         Allocated memory       4 kB         Dataset descri       Column percent contribution         3 attribute(s)       Presiduals         205 example(s)       Contribution thresold         Attribute       Category	
		C	
Data visualization	Statisti	ics Nonparamet OK Cancel He	lp
Feature selection	Kegress		
opvitearning	/weta-spv le	arning opvitearning assessment Scoring Association	
Categorical r	🖾 Goodma	an-Kruskal Lambda 🔄 Kendall's tau 👜 Mood Runs Test 🔄 Spear	man's rho
Friedman's ANOVA by R	nks 💾 Kendall	sinkruskai iau 🔐 kruskai waitis i-way an OVA 🖬 Partial Inetto 🗾 El Inett 's Concordance W 📊 Mann-Whitney Comparison 🕂 Sien Test	U Wolfowitz
	Internation		
<			>
			:

We click on the VIEW menu.



The expected frequencies are now printed under the observed frequencies.

We can also obtain the cells contributions. They are useful in determining which ranges of the two variables depicted in a table are contributing the most to the overall relationship. We click again on the PARAMETERS menu and select, in the ANALYSIS tab, the Chi-square contributions option.

🛣 TANAGRA 1.4.18 - [Dataset (tan2	3C.txt)]	<b>- - ×</b>
Tile Diagram Component Window I	telp.	- 8 ×
	Cross-tabulation parameters	
Analysis		×
🖃 🏢 Dataset (tan23C.txt)	Parameters Analysis	
🖮 🚰 Define status 1		
🔚 Contingency Chi-Square	1 Additional information	
	ONone	
	O Row percent	=
	O Column percent	
	Chi-square contribution	
	O Expected values	
	O Residuals	
	Contribution thresold Highlight contribution greater than 2.00 × average	<b>v</b>
<u>با</u>		
Data visualization St	OK Cancel Help	ature construction
Feature selection Res	ression Factorial analysis PLS	Clustering
Spv learning Meta-s	pv learning Spv learning assessment Scoring	Association
🖾 Categorical r 🛛 🕮 Go	odman-Kruskal Lambda 🛛 🖄 Kendall's tau 🛛 🛋 Mood Runs Test	💒 Spearman's rf
🔚 Contingency Chi-Square 🛛 🗾 Go	odman-Kruskal Tau 🔣 Kruskal-Wallis 1-way ANOVA 🛛 🖪 Partial Theil U	🖾 Theil U
Friedman's ANOVA by Ranks 👫 Kei	ıdall's Concordance W 🛛 🚻 Mann-Whitney Comparison 🛛 📑 Sign Test	📺 Wald-Wolfowi
<		>

We obtain the following results.

TANAGRA 1.4.18 - [Contingenc	y Chi-S	quare 1]								
Tile Diagram Component Window	Help								-	a ×
Analysis					Results					
■ I Dataset (tan23C.txt) ■ T Define status 1		Row (Y	) Column (X)	imn Statistical indicator			Cross-tab			
🔛 🔛 Contingency Chi-Squa	re 1			Stat	Value		gas	diesel	Sum	
				Tschuprow's t	0.164399	high	41	0	41	
				Cramer's v	0.164399		(+ 8 %)	(- 72 %)		
				Phi²	0.027027	low	144	20	164	
		Consumpt	tion fuel-type	Chi² (p-value)	5.54 (0.0186)	Sum	183	20	205	=
				Lambda	0.000000			V	100/8	
						Tau (p-value)	0.0270 (0.0189)			•
				U(R/C) (p-value)	0.0461 (0.0021)					
										~
			Co	omponents						
Data visualization	Statistic	25	Nonparame	tric statistics	Instance	selectio	n F	eature cons	truction	
Feature selection R	egressio.	on	Factoria	al analysis	PI	.s		Cluster	ing	
Spv learning Meta	spv lea	arning	Spv learnin;	g assessment	Sco	ring		Associat	ion:	
🖉 Categorical r 📃 G	ioodmar	n-Kruskal La	mbda 🔣 🕅	(endall's tau	1	⊾ Mood	Runs Test		🕍 Spearma	an's rt
🖬 Contingency Chi-Square 🛛 🖾 G	ioodmar	n-Kruskal Ta	u 📶 K	(ruskal-Wallis 1-w	/ay ANOVA 🚦	🗿 Partial	l Theil U		🗾 Theil U	
📰 Friedman's ANOVA by Ranks 斗 K	endall's	Concordan	ce W 📶 📶	Aann-Whitney Co	omparison 🚊	i Sign T	est		📥 Wald-Wo	olfowi
<										>
										:

The contributions are displayed in percent of the whole chi-square statistic (CHI-2 = 5.54 in our dataset). When the contribution is 2 times (we can modify this) larger than the average contribution, it is underlined.

In our table, we observe that the relation between CONSUMPTION and FUEL-TYPE relies mainly on the opposition (-72%) between diesel fuel type and high consumption.

Of course, this kind of analysis is only interesting if the CHI-2 is significant.

### Asymmetric association

In the most of cases, the variables have not a symmetrical role in the analysis. We want to check if one variable determines the values of another variable. For instance, we want to know if the FUEL-TYPE determines the CONSUMPTION. A first way to check this kind of association is to display the column percentage in the table. Then we can determine if the proportion of cars with low (high) consumption are different according to their fuel type. To do that, we modify again the parameters of the component.

TANAGRA 1.4.18 - [Dataset (tan23C.t)	t)]	
Tile Diagram Component Window Help		_ 8 ×
Analysis	Dataset (tan23C.txt)	
🖃 🎹 Dataset (tan23C.txt)	Parameters	
🖃 🏠 Define status 1		
Contingency Chi-Square 1	Cross-tabulation parameters	
	Parameters Analysis	
	Additional information None Row percent Column percent Chi-square contribution Expected values Residuals Contribution thresold	~
Data visualization Statistic	Highlight contribution greater than 2.00 × average	ture construction
Feature selection Regression		Clustering
Spvilearning Meta-spvilea		Association
🔄 Categorical r 🛛 🗐 Goodman	OK Cancel Help	🖾 Spearman's rf
🔛 Contingency Chi-Square 🛛 🗾 Goodmar	-Kruskal Tau 📶 Kruskal-Wallis 1-way ANOVA 🖪 Partial Theil U	🖾 Theil U
Friedman's ANOVA by Ranks 👫 Kendall's	Concordance W 🛛 🗽 Mann-Whitney Comparison 📑 Sign Test	📥 Wald-Wolfowi
	ш )	

We obtain the following table.



We note that the consumption is different according to whether the car uses diesel or gas. Working on percentages makes the columns comparable, which authorizes this kind of analysis.

#### PRE measures (Proportionate Reduction in Error)

PRE measures are asymmetric. They allow characterizing a prediction analysis (TARGET vs. INPUT). Three measures are available: Goodman & Kruskal's Lambda; Goodman & Kruskal's Tau; Theil's U (Uncertainty Coefficient). The last two indicators display also the p-value of the significance test.

It is possible to obtain the details of calculations using specific components.

**Goodman-Kruskal Tau.** We add the GOODMAN-KRUSKAL Tau component (NONPARAMETRIC tab) into the diagram, under DEFINE STATUS 1. We obtain the following results when clicking on the VIEW menu.

💯 TANAGRA 1.4.18 - [Go	oodman-Kruskal Ta	u 1]							
🕎 File Diagram Componen	t Window Help							- 🖻 🗙	
D 📽 🖪  🏭									
Analysis									
🗉 🥅 Dataset (tan23C.txt	:)			Goodman-Kr	uskal Tau 1				
🛓 🎦 🙀 Define status 1				Paramo	eters				
Contingenc	y Chi-Square 1	Doculte							
🔤 🗹 🔤	ruskal Tau 1			NO3C					
1		Goodmar	n & Kruskal	's Tau fo	or nomi	nal attr	ributes		
		Y	Х	Tau	Chi <sup>2</sup>	d.f.	p-value		
		Consumption	fuel-type	0.027027	5.51	1	0.0189		
		Computation tin	ne : 0 ms.						
			5572007 10.25.51						
			Components						
Data visualization	Statistics	No	nparametric stat	istics I	nstance sele	ection			
Feature construction	Feature select	ion	Regression		Factorial an	alysis			
PLS	Clustering		Spv learning		Meta-spv lea	irning			
Spv learning assessment	Scoring		Association						
🔄 Categorical r	Friedman's	ANOVA by Ran	🗧 🗉 Goodman	-Kruskal Tau	> 🖄	endall's tau		Mai Mai	
🖽 Contingency Chi-Square	e 🔲 Goodman-K	ruskal Lambda	🕴 Kendali's	Concordance	e W 👘 📊 K	ruskal-Walli:	s 1-way ANOVA	A 🛝 Mo	
<								>	
L									

Tau statistic between CONSUMPTION (TARGET) and FUEL-TYPE (INPUT) is displayed again (0,0270).

The computed chi-square is a transformation of Tau (see Light & Margolin, 1971). It is 5.51. The degree of freedom is the same than the test of independence. The obtained p-value (0.0189) enables to conclude to a significant association at the 5%-level.

Theil's U. We add now the THEIL U component into the diagram.



The Theil's U is 0,046 in our analysis. There is also a chi-square transformation of this statistic, says also "Likelihood Ratio Chi-Square" statistic in some software. Then we can compute the *p*-value of the test. We observe the same conclusion as the preceding test.

For comparison, STATISTICA gives the following results.

🔚 Stats : CONSUMPT(2) x FUEL_	TYP(2) (fuel_co	- 🗆 ×	
STAT. ELEMENT.	Chi²	dl	р
Chi² de Pearson	5.540541	df=1	p=.01858
Chi² du MV	9.452478	df=1	p=.00211
Coeff. d'incertitude	¥=.0460726	¥=.0721161	X Y=.05622

For the Theil's statistic, TANAGRA computes also the asymptotic variance/standard-deviation. Then, to give an idea of the variability of the solution, the 95% confidence interval is presented.

### Partial association

In some situations, we may seek to control for the effects of a variable E on the association between two nominal variables A and B.

For more details about this problem, see the following websites:

- http://www2.chass.ncsu.edu/garson/PA765/association.htm#spss
- http://www2.chass.ncsu.edu/garson/pa765/partialr.htm

In our dataset, ASPIRATION is the third variable that can play an important role (standard or turbo).

**Partial Theil's U.** We add a DEFINE STATUS component at the node of the diagram. We set CONSUMPTION as TARGET, FUEL-TYPE as INPUT, and ASPIRATION as ILLUSTRATIVE.



Then, we add the PARTIAL THEIL U component (NONPARAMETRIC STATISTICS tab).



We observe that the partial association U(AB/E) = 0.0876 is larger than the association U(AB) = 0.0460. It seems that the association between CONSUMPTION and FULE-TYPE is more significant for ASPIRATION=TURBO cars. Indeed, we see in the table that the conditional association U(AB/E=turbo)=0.3099. TANAGRA computes also the asymptotic standard deviation and the 95% confidence interval.

The partial association is a weighted average of the conditional association index. The weight depends on the frequencies of the control value but also on the prediction error. See Olszak and Ritschard (1995) for detailed explanation<sup>2</sup>.

<sup>&</sup>lt;sup>2</sup> M. OLSZAK, G. RITSCHARD (1995) « The behavior of nominal and ordinal partial association measure », in The Statistician, vol.44, n°2, pp.195-212.

**Conditional association.** We want to lead a deeper analysis on the subset of cars ASPIRATION=TURBO. We must apply all the computations on the examples corresponding to this characteristic.

To do that, we add the DISCRETE SELECT EXAMPLES component (INSTANCE SELECTION tab) into the root of the diagram. We set the attribute selection to ASPIRATION and the value selection to TURBO.

TANAGRA 1.4.18 - [Partial Theil U 1]		
Arelysis  The Dataset (tan23C.txt)  Define status 1  Contingency Chi-Square 1  Contingency Chi-Square 1  Theil U 1  Theil U 1  Contine status 2  Figure 1 a Theil U 1  Concrete select examples 1  Concrete select examples 1	Results Partial Uncertainty Coefficient Attribute-value examples selection Parameters Attribute : aspiration Yalue : turbo Value : turbo	
Data vitualization Statisti Feature election Regress Spulearning Meta-spule Continuous telect examples & Rule-ba Discrete select examples Sampling Recover examples Stratifie	ed selection d sampling	

Then, we insert again the DEFINE STATUS and CONTINGENCY CHI-SQUARE components. TARGET attribute is CONSUMPTION, and the INPUT one is FUEL-TYPE. We obtain the following results when we click on the VIEW menu.



There are 37 ASPIRATION=TURBO cars in the database. All the FUEL-TYPE=DIESEL cars present a low CONSUMPTION. At the opposite, a majority (54.17%) of FUEL-TYPE=GAS cars have a high consumption. The association is significant at the 1% level (p-value = 0.0001).

#### Variables interaction

Another kind of analysis is to analyze the interactions between FUEL-TYPE and ASPIRATION in the determination of CONSUMPTION. The easiest way is to create a new variable as follows.

**Creation of a new variable.** We close TANAGRA. In the EXCEL spreadsheet, we add a new column (FUEL-ASPIRATION). We insert the following formula<sup>3</sup>.

🖾 м	icrosoft Exce	l - fuel_cons	umption.xls					
	<u>Fi</u> chier <u>E</u> dition	Affichage Inse	rtion Forma <u>t (</u>	<u>)</u> utils <u>D</u> onnées Fe	e <u>n</u> être <u>?</u> Tana	agra Sipina	_	BX
1 D	🚔 🔲 🚑 [	a 🤒 🖁	a 🙉 🛷 🗠	- α - 🥘 . Σ	: f* 🗟 🌗	Z1 🛍 🛺	100% - »	🕭 - 2
1 -	D2 -		ONCATENER(E	32;" ";C2)			• ] •	
	A	В	с	D	E	F	G	=
1	Consumption	fuel-type	aspiration	fuel-aspiration				<b>_</b>
2	high	gas	turbo	gas turbo				
3	high	gas	turbo	gas turbo				
4	high	gas	std	gas_std				
5	high	gas	std	gas_std				
6	high	gas	std	gas_std				
7	high	gas	std	gas_std				
8	high	gas	std	gas_std				
9	high	gas	turbo	gas_turbo				
10	high	gas	turbo	gas_turbo				
11	high	gas	std	gas_std				
12	high	gas	turbo	gas_turbo				
13	high	gas	std	gas_std				
14	high	gas	std	gas_std				
15	high	gas	std	gas_std				
16	high	gas	std	gas_std				
17	high	gas	std	gas_std				
18	high	gas	std	gas_std				
19	high	gas	turbo	gas_turbo				
20	high	gas	turbo	gas_turbo				
21	high	gas	std	gas std				
	▶ ▶ \datas	et /			•			
De <u>s</u>	sin + 🗟 🌀	Formes automal	tiques 🗸 🔪 🍾		. 🙍 🔺	🖉 - 🛕 - 💻	🛲 🛱 🗖 🎒	-
Prêt							NUM	

*New diagram.* We select the cells range and click on the TANAGRA/EXECUTE TANAGRA menu.

🛛 м	licrosoft Exce	el - fuel_con	sumption.xls					-			
8	Eichier Edition	Affichage Ins	ertion Forma <u>t</u>	<u>O</u> utils <u>D</u> onnées F	enêtre	<u>?</u> Tana	agra Sipina		_ 8 ×		
D	28	🗟 🖤 🐰	à 🛍 🝼 🖡	n + 04 + 🍓	Σ f*	5€ 2↓	XI 🛍 4	100% - 😤	🕭 - "		
1		= =	CONCATENER(	B206;"_";C206)							
	A	B	С	D		E	F	G			
1	Consumption	fuel-type	aspiration	fuel-aspiration					-		
2	high	gas	turbo	gas_turbo	1						
3	high	gas	turbo	gas_turbo							
4	high	gas	std	gas_std	3			1			
5	high	gas	std	gas_std	1						
6	high	gas	std	gas_std							
7	high 👝	ase	etd	hts sen	100						
8	high F	Execute Tanagra									
9	high 🦰										
10	high										
11	high	Dataset r	ange (including tr	he name or the attri	butes	- rirst row,	<u>к</u>				
12	high	\$A\$1:\$D\$206									
13	high			<u> </u>		-					
14	high			OK			Cancel				
15	high						cancor				
16	high										
17	high	gas	std	gas_std		1					
18	high	gas	std	gas_std	T.						
19	high	gas	turbo	gas_turbo							
20	high	gas	turbo	gas_turbo	3						
21	high	gas	std	gas std	1						
4	► ► \ datas	et /				4			•		
Des	sin + 🔓 🕝	Formes autom	atiques 🕶 🔨 🖄			A - :	<u>⊿</u> - <u>▲</u> - ≡	= = = (	7.		
Prêt								NUM			

 $<sup>^{3}</sup>$  In English EXCEL version, we use the following function =CONCATENATE(...).

We insert the DEFINE STATUS component. We set CONSUMPTION as TARGET, FUEL-ASPIRATION as INPUT.

	Define attribute statuses			
TANAGRA 1.4.18 - [Dataset (tan252.txt)]				
Tile Diagram Component Window Help	Parameters	_ 8 ×		
	Attributes : Target Target Strative			
Analusis	Consumption Consumption			
Ai laiyoto	D aspiration	txt) 🍧		
Dataset (tan252.txt)	D fuel-aspiration			
Define status 1		in252.txt		
	E E Clear selection			
		-		
	OK Cancel Help			
	Define attribute statuses			
	Parameters			
	Attributes : Target Input			
	Consumption     fuel-aspiration	×		
Data visualization Statistics	L fuel-aspiration	Feature construction		
Feature selection Regression		Clustering		
Conclosed of the state of the s		Association		
Spvilearning ///leta-spvilearning		Association		
Correlation scatterplot				
Export dataset 🛛 🕍 Scatterplot with				
	E E XX Clear selection			
	OK Cancel Help			

We add the CONTINGENCY CHI-SQUARE component. We select the Chi-Square contributions option.

T	TANAGRA 1.4.18 - [Co	ontingency Chi-	Square 1]											
Ť	File Diagram Componen	t Window Help										-	Ξ×	
	Analysis		Additional i	information		3							^	
🖃 🏢 Dataset (tan252.txt)			Contributio	on thresold		2.0								
	😑 🚰 Define status 1													
	Contingenc	y Chi-Square 1	Results											
			Row (Y)	Row (Y) Column (X) Statistical indicator			Cross-tab							
					Stat	Value		gas_turbo	gas_std	diesel_turbo	diesel_std	Sum		
				Tschuprow's	t 0.255589	high	13	28	0	0	41			
			Consumption	on fuel-aspiration	Cramer's v	0.336374	-	(+ 60 %)	(- 2 %)	(- 11 %)	(-6%)			
	/	Phi <sup>z</sup>			0.113147	low	11 (- 15 %)	133 (+ 1 %)	13 (+ 3 %)	(+ 2 %)	164	=		
	/	Chi² (p-valu			e) 23.20 (0.0000)		24	161	13	7	205			
					Lambda	0.048780	Sum					100%		
					Tau (p-value	e) 0.1131 (0.0000)								
					U(R/C) (p-va	alue) 0.1135 (0.0000)								
<		>											~	
					C	omponents								
	Data visualization Statistic			Nonparametric	Instance selection		Feature construction			Feature selection				
	Regression	Factorial a	nalysis	PLS		Clustering		Spv learning			Meta-spv learning			
Sp	Spy learning assessment Scoring Assoc					tion								
📴 🕻 teeorical r 🔠 Friedman's ANOVA by Ranks 🗵 Goodman-Kruskal Tau 🖉 Kendall's tau 🕍 Mann-Whitney Comparison								🛱 Partial	Theil					
	Contingency Chi-Square	e 🚬 🗏 Goodma	an-Kruskal Lar	mbda 👫 Keno	dall's Concor	dance W 👘 📊 Kru	skal-W	allis 1-way AN	NOVA 🗼 A	Aood Runs Te	st	📑 Sign Te	est	
											2			

The CHI-QUARE statistic is 23.20. The relation is very significant (p-value < 0.0000...). The main information (60%) into the contingency table is the "high" consumption of TURBO-GAS cars.

## Conclusion

CHI-SQUARE statistic is widely used in order to gauge the relation between two nominal variables. In this tutorial, we have shown how to implement this kind of analysis with TANAGRA.

In addition, because the chi-square and derived measures are symmetric, they cannot handle a prediction/explanation process, we present also some asymmetric measures.