Subject

Descriptive statistics with TANAGRA.

The aim of descriptive statistics is to describe the main features of a collection of data in quantitative terms¹. The visualization of the whole data table is seldom useful. It is preferable to summarize the characteristics of the data with some selected numerical indicators.

In this tutorial, we distinguish two kinds of descriptive approaches: the univariate tools which summarize the characteristics of a variable individually; the bivariate tools which characterize the association between two variables. According to the type of the variables (categorical or continuous), we use different indicators.

Creating a diagram

Dataset

The dataset comes from a consumer survey about a product for baby. There are 20 observations, ID is an identifier. The variables are interval (REVENU – INCOME), categorical (VILLE – TOWN) or ordinal (OPINION – JUDGEMENT). The number of children is more complicated. It is both a continuous variable (the difference between two values has a signification) and an ordinal variable (the number of values is limited).

🖾 м	icrosoft Exce	l - enquete_s	atisfaction_f	emmes_1953	l.xls			
	<u>Fi</u> chier <u>E</u> dition	Affichage Inse	rtion Forma <u>t</u> (utils <u>D</u> onnées	Fe <u>n</u> être <u>?</u> Ta	nagra Sipina		_ & ×
l n	🗠 🗖 🖾 I	אַייא אַ <u>אַ</u>	a 🗈 🛷 🔽	1 + CI + 🙆	Σ. £. 🗟 4	1 ZI 10 🕡	100% - 8] ,00 🕭 , »
					~)~ WEI 2		· · · · · ·	₽ ▼]+,0 <mark>-*</mark> . ▼
	A1 _	B	C	D	F	F	G	н
1	Id	NbEnfants	Revenu	Ville	Opinion		0	
2	1	4	30000	Paris	Mauvaise			
3	2	0	60000	Montpellier	Très bonne			
4	3	1	9000	Rouen	Mauvaise			
5	4	1	15000	Paris	Mauvaise			
6	5	2	40000	Marseille	Moyenne			
7	6	2	40000	Marseille	Moyenne			
8	7	2	10000	Nice	Bonne			
9	8	3	45000	Paris	Moyenne			
10	9	3	80000	Nancy	Passable			
11	10	4	50000	Nice	Moyenne			
12	11	2	60000	Nice	Passable			
13	12	3	55000	Marseille	Bonne			
14	13	4	85000	Montpellier	Bonne			
15	14	5	60000	Nice	Moyenne			
16	15	2	40000	Rouen	Bonne			
17	16	1	10000	Nancy	Passable			
18	17	3	120000	Nice	Mauvaise			
19	18	3	80000	Paris	Moyenne			
20	19	4	18000	Paris	Tres bonne			
21	20		90000	Marseille	Passable	<u> </u>		
	N N dame	has /						
		ees/						
Des	sin 🕶 🔓 🌀	Formes automal	tiques 🕶 🔪 🍾		4 🙎 🖄 -	<u>- 4</u> - <u>A</u> - <u>E</u>	■ ☴ ☴ ◘	•
Prêt					Somme=9	997264	NUM	

Importing the data file

First, we must create a new diagram and import the data file « enquete_satisfaction_femmes_1953.xls ». The easiest way is to open the file in Excel. We select the range of cells, including the first row corresponding to the name of the variables, and we click on the TANAGRA / EXECUTE TANAGRA menu².

¹ <u>http://en.wikipedia.org/wiki/Descriptive_statistics</u>

	licrosoft Exce	l - enquete_s	satisfaction_f	emmes_195	3.xls	K			×
	Fichier Edition	Affichage Inse	rtion Forma <u>t</u> (<u>Dutils</u> Données	Fenêtre ? Tar	nagra Sipina		_르.	×
		<u>}</u> ₩ ₩	b 🛍 🝼 🕑) + (4 + 🔮	$\Sigma f_* = \frac{1}{2}$		🚜 100% 👻 🖸	ये - २% 🥭 -	*
		= Id							
	A	В	С	D	E	F	G	H ·	-
1	ld	NbEnfants	Revenu	Ville	Opinion				
2	1	4	30000	Paris	Mauvaise				
3	2	0	60000	Montpellier	Très bonne				
4	3	1	9000	Rouen	Mauvaise				
5	4	1	15000	Paris	Mauvaise				
6	5		40000					· ·	
7	6	Execute Ta	nagra				×		
8	7	-							
9	8	Datase	at range (includin	g the name of th	ne attributes firs	st row):			
10	9	\$A\$	1:\$E\$21				-		
11	10								
12	11	-			ок	Cancel			
13	12	-							
14	13				1				
15	14	5	60000	Nice	Moyenne				
16	15	<u>∠</u>	40000	Rouen	Bonne	7			
10	10	2	10000	Nancy	Passable		_		
10	1/	3	120000	Daria	Movenne				
20	10	1	18000	Daria	Tràc honno	-			
20	19	- 4	18000	Marcoillo	Baccable	-			
21	20	J	90000	Marselle	rassable				
	N N donné	inc /		l			1		-
Des		Formes automal	tiques 🗸 🔪		4 2 3 -	.1 . A .			
Prêt		_			Somme=9	97264	NUM		

TANAGRA is automatically launched. A new diagram is created and the dataset is available for the analysis.

🌋 TANAGRA 1.4.19 - [Da	taset (tan22.txt)]									
T File Diagram Component	Window Help				_ 8 ×					
🗅 📽 🔜 🔯										
	Analysis				~					
Dataset (tan22.txt)	4			Dataset (tan22.txt)						
			Database (COD)	Parameters	n)ten 22 test					
			DataDase : C:DC	COWE~ UMBISOLUTIOCALS~ LUGH	pitalizz.txt					
				Results	=					
			Download	l information						
			Datasource pro	cessing						
			Computation tim	ne Oms						
			Allocated memor	y 6 KB						
	Computation time 0 ms Allocated memory 6 KB Dataset description									
			5 attribute(s)							
			20 example(s)		~					
			Components							
Data visualization	Statistics	Nonpara	ametric statistics	Instance selection	Feature construction					
, Feature selection	Regression	Fact	torial analysis	PLS	Clustering					
Spy learning	Meta-spy learning	Spv lear	rning assessment	Scorine	Association					
	1.8 Carata and a control of the			Ū.						
Export dataset	Wiew datacet									
Scatterplot	View multiple scatter	alot								
L										

Analyzing Categorical Variables

Univariate statistics

The first tool for characterizing the distribution of a categorical attribute is the absolute frequency. For each value of the variable, we count the number of examples. We can also compute the percentage of examples.

² This menu is available after we install the TANAGRA.XLA add-on into Excel. The procedure is described here: <u>http://data-mining-tutorials.blogspot.com/2008/10/excel-file-handling-using-add-in.html</u>

We insert the DEFINE STATUS component into the diagram using the shortcut into the tool bar. We set VILLE and OPINION as INPUT attributes.

TANAGRA 1.4.19 - [Dat	aset (tan22.txt)]		<
File Dagran Component	Window Help		×
Data visualization Feature selection Spy learning Correlation scatterplot Export dataset Scatterplot	Statist Regres Meta-spv I Soatterplot with label View dataset L'View multiple scatterplot	uction	

We add now the UNIVARIATE DISCRETE STAT (STATISTICS tab) into the diagram. We click on the VIEW menu in order to obtain the results.



We note for instance that 25% of the respondents come from Paris; 10% from Montpellier; etc.

The Gini index helps to locate the concentration of observations on some values. If it is close to 0, the observations are concentrated on a few values of the variable.

Bivariate descriptive statistics for categorical variables

We want to characterize now the distribution of OPINION according to VILLE. We create a contingency table.

Absolute frequencies

After the DEFINE STATUS component into the diagram, we add the component CONTINGENCY CHI-SQUARE (NONPARAMETRIC STATISTICS tab). We click on the PARAMETERS contextual menu. The variables are listed in the INPUT list i.e. the component computes the contingency table for each pair of variables.

TANAGRA 1.4.19 - [Univariate discrete stat 1 File Diagram Component Window Help	1			
Analysis Analysis C Dataset (tan22.txt) C tan22.txt 1 C ta	Cross-tabulation param	eters s		
Contingency Chi-Square 1	Sort results Sort by Row attribute n	ame		-
	Input list • Cross Input	e name istic		ogram
		OK Cance	I Help	×
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	
Categorical / 2 Goodman-Kruska Contingency Chi-Square 2 Goodman-Kruska Friedman's ANOVA by Ranks III Kendalls Concor	l Lambda 🖄 Kendall's tau I Tau 🚛 Kruskal-Wallis 1- dance W 🛄 Mann-Whitney C	a Mood Runs way ANOVA I Partial Theil omparison ∷ Sign Test	Test 🔄 Spearr . U 🗐 Theil I 📥 Wald-V	nan's rho U ∦olfowitz Runs Test ▶

We click on the VIEW menu, we obtain the following results.



We observe for instance that two individuals from Paris have a bad (mauvaise) opinion about the product; 1 individual from Nice has a good (bonne) opinion; etc.

Various numerical indicators try to quantify the strength of the association between the variables (see http://www2.chass.ncsu.edu/garson/PA765/assocnominal.htm or http://en.wikipedia.org/wiki/Crosstab).

Percentage by row or percentage by column

Often, the absolute frequencies are not informative. It is better to compute relative frequencies, by row or by column. About the opinion according the town, it is more suitable to dividing the absolute values by the margin. The values are then comparable.

Cross-tabulation parameters
Parameters Analysis
Additional information None Row percent Column percent Chi-square contribution Expected values Residuals
Contribution thresold Highlight contribution greater than 2.00 × average
OK Cancel Help

We click on the PARAMETERS menu. We select the ROW PERCENT option.

By clicking on the VIEW menu, we obtain:

TANAGRA 1.4.19 - [Contingency Chi-Squa	ire 1]											
											_	
Analysis	Row (Y)	Column (X)	Statistical	indicator				Cross-tab				^
Define status 1			Stat	Value		Mauvaise	Très bonne	Moyenne	Bonne	Passable	Sum	
III Univariate discrete stat 1 Contingency Chi-Square 1			Tschuprow's t Cramer's v	0.503409 0.532291	Paris	2	20.00%	2	0	0	5	
			Phi ^z Chi ^z (p-value)	1.133333 22.67 (0.3055)	Montpellier	40.00%	20.00%	40.00%	0.00%	0.00%	2	
			Lambda	0.200000	Rouen	1	0	0.00%	1	0.00%	2	
	Ville	Opinion	Tau (p-value)	0.1925 (0.5682)	Marseille	50.00%	0.00%	2	50.00%	0.00%	4	
			U(R/C) (p-value)	0.3598 (0.2192)	Nice	0.00%	0.00%	50.00%	25.00% 1	25.00%	100%	
					Nancy	20.00%	0.00%	40.00%	20.00%	20.00%	100% 2	
					Sum	4	2 10%	6 30%	0.00% 4 20%	4 20%	20	
						2070	1070	3078	2070	2070	1007	~
	1.000			Component	8							
Data visualization Statistics		Nonparamet	ric statistics	Instance se	lection	Feature o	construction	n Fe	ature selec	tion		
Regression Factorial analysi Spy learning assessment Scoring	Factorial analysis PLS			Cluster	ing	Spv	learning	///	eta-spv leari	ning		
Categorical r Goodman-Kru Goodman-Kru Goodman-Kru Goodman-Kru Friedman's ANOVA by Ranks	uskal Lamb uskal Tau icordance	oda 🖾 K Mil K W Mil W	Kendall's tau Mood Mruskal-Wallis 1-way ANOVA Partial Mann-Whitney Comparison Sign Ti			Test U	🕍 Spe 🖭 The	arman's rho il U J-Wolfowitz	Runs Test	::#Wilcox	on Signed Rar	ıks T
)			>

For instance, the absolute frequency of good opinion is the same in Montpellier and Nice. When we divide the value by the number of instances in each town, we note that the proportions are not the same.

Continuous variables

Univariate statistics

Two components are dedicated to univariate descriptive statistics for continuous variable.

We insert again the DEFINE STATUS component. We set as INPUT both NBENFANTS (number of children) and REVENU. Computing descriptive statistics on the ID column is not really useful.

TANAGRA 1.4.19 - [Continger	ncy Chi-Square 1]									
Analysis	Row (Y)	Column (X)	Statistical indicator		(Cross-tab				
Define status 1	Define	attribute st	atuses ,			loyenne	Bonne	Passable	Sum	
Contingency Chi-Sq	uare 1 Par	ameters			_	2 40.00%	0 0.00%	0 0.00%	5 100%	
	ŀ	uttributes :		Target Input Illustrative		0 0.00%	1 50.00%	0 0.00%	2 100%	
		C Id NbEnfants		NbEnfants 4		0 0.00%	1 50.00%	0 0.00%	2 100%	
		C Revenu				2 50.00%	1 25.00%	1 25.00%	4 100%	
		U Opinion	•			2 40.00%	1 20.00%	1 20.00%	5 100%	
						0 0.00%	0 0.00%	2 100.00%	2 100%	
						6 30%	4 20%	4 20%	20 100%	
		8, 8,	跌	Clear selection					4	
Data visualization	Statistics					Fe	ature selec eta-sov lear	tion		
w learning assessment	Scoring	HSSUCIAL			ip		.cu 5pr (641)			
Categorical r 🛛 🗐 Contingency Chi-Square 🗐 Friedman's ANOVA by Ranks 👫	Goodman-Kruskal Lamb Goodman-Kruskal Tau Kendall's Concordance	da 🖾 Ken 🙀 Krus W 🎪 Man	dall's tau skal-Wallis 1-way ANOVA in-Whitney Comparison	Mood Runs Test Mood Runs Test Martial Theil U Partial Theil U Sign Test	Spear Theil Wald-1	man's rho U Wolfowitz F	uns Test	⊒ [#] Wilcoxo	n Signed Ra	ank

Then we add the UNIVARIATE CONTNUOUS STAT (STATISTICS tab) component. We click on VIEW menu.

💇 TANAGRA 1.4.19 - [Univ	ariate continuous stat	1]								
Tile Diagram Component	Window Help								- 8 ×	
🗅 🚅 🔲 🎎										
Analysis		Univariate continuous stat 1								
🖃 🏢 Dataset (tan22.txt)							Parameters			
😑 🛟 Define status 1	Att	ributes	:2							
- III Univariate disc	rete stat 1 Exa bi-Square 1 -	Examples : 20								
🖃 🚰 Define status 2							Results			
🔤 📶 Univariate con	tinuous stat 1 🛛 🗛	tribute	Min	Max	Average	Std-dev	Std-dev/avg	J		
	Nb	Enfants	0	5	2.7000	1.3803	0.5112	}		
	Re	venu	9000	120000	49850.0000	30525.7010	0.6124	J		
	Con	nputatio ated at	n time 19/06	: 0 ms. /2007 13	3:41:40					
					Component	s				
Data visualization	Statistics	Nor	parar	netric s	tatistics	Instanc	e selection	Feature construction		
Feature selection	Regression		Facto	orial ana	lysis		PLS	Clustering		
Spv learning	Meta-spv learning	Spv	lear	ing ass	essment	So	coring	Association		
🕁 Bartlett's test	🚮 Group exploration		۸	lormalit	y Test	Linter,	T-Test			
A Brown - Forsythe's test	⊷ Levene's test			Dne-way	ANOVA		T-Test Unequ	al Variance		
Group obspatarization	Linear correlation	nt ct st	- <u> ≪</u> (- + ‡ r	Jne-way Deirod T	Tost	- <u>-</u>	Univariate co	ntinuous stat		
	Pismore Universate Cu	ni stat	<u>+1</u>	aneu I	-1650		onneanate us	crete stat		

This component computes simple indicators such as mean, standard deviation, etc. It is useful when we want to obtain quickly an overview on the characteristics of a large number of variables.

A second tool gives more detailed information. We add the MORE UNIVARIATE CONTNUOUS STAT component (STATISTICS tab). We obtain the following results.

¥ 🖬 🙀										
Analysis	Attribute	Sta	ats			Histogram				
Dataset (tan22.txt)		Stati	stics	Values	Count	Percent	Histogram			
🖹 🛟 Define status 1		Average	2.7000	x_<_0.5000	1	5.00%				
Univariate discrete stat 1		Median	3.0000	0.5000_=<_x_<_1.0000	0	0.00%				
E Contingency Chi-Square 1		Std dev. [Coef of variation]	1.3803 [0.5112]	1.0000 =< x < 1.5000	3	15.00%				
上述 Univariate continuous stat 1		MAD [MAD/STDDEV]	1.1300 (0.8187)	1 5000 - < × < 2 0000	0	0.00%				
More Univariate cont stat 1		Min*Max (Full range)	0.00 * 5.00 [5.00]	1.5000_44_4_4_2.0000		0.00%	_			
1	NbEnfants	1st*3rd quartile [Range]	2.00 * 4.00 [2.00]	2.0000_=<_x_<_2.5000	5	25.00%	<u> </u>			
/		Skewness (std-dev)	-0.0640 (0.5121)	2.5000_=<_x_<_3.0000	0	0.00%	_ (
		Kurtosis (std-dev)	-0.6171 (0.9924)	3.0000_=<_X_<_3.5000	5	25.00%				
				3.5000_=<_X_<_4.0000	0	0.00%				
				4.0000_=<_x_<_4.5000	4	20.00%				
				x>=_4.5000	2	10.00%	<u> </u>			
	<			ly III						
			Components							
Data visualization Statistics	Nor	parametric statistics	Instance selection	Feature construction	Feature	e selection	Regression			
actorial analytis PLS Association		Clustering	Spv learning	Meta-spv learning	Spv learnir	ng assessme	nt Scoring			
artlett's test	ration	A Normality Test	La T-Test							
rown - Forsytheis test Albeen corre	: lation	ML One-way ANOVA	way ANOVA La T-Test Unequal Variance							
	the seat state	L Daired T.Tert	En universate continuous stat							

Various indicators and additional information about the shape of the distribution are supplied.

Bivariate statistics for continuous variables

Scatter plot

Unlike categorical variables, some graphical tools allow to evaluate the association between continuous variables. We insert the SCATTERPLOT component (VISUALIZATION tab). We obtain a scatter plot. We can modify interactively the variables in the horizontal and the vertical axis. The size of the points can be modified.



Numerical indicators

We can characterize also the association with some numerical indicators. It is especially useful when we deal with a large number of variables. For instance, the LINEAR CORRELATION (STATISTICS tab) component allows to compute the correlation coefficient between the variables. In the settings dialog box, we indicate that the variables are listed in the INPUT section of the preceding DEFINE STATUS component.

TANAGRA 1.4.19 - [Scatterplot 1]		_ 🗆 🗙
Tile Diagram Component Window Help		- 8 ×
🗅 🛩 🔜 👫	relation options	
Analysis	Parameters Sort results Sort by Y attribute name X attribute name r - value Inj- value Injut list Target and Input C Torss Input	120 000
Data visualization Statistics Feature selection Regression Spv learning Meta-spv learning Bartlett's test Group exploration Fisher's test Learnactest Group characterization More Univariate cont stat	OK Cancel Help	

We obtain the following result.

Image: Components Components Components Components Components Components Components Components Components Constraining assessment Social analysis Data visualization Statistics Nonparametric statistics Instance selection Feature construction Spy learning Meta-spy learning Spy learning Spy learning Spy learning Spy learning Spy learning Spy learning Spy learning Spy learning Spy learning Spy learning Spy learning Spy learning Spy learning Spy l	TANAGRA 1.4.19 - [Lin	near correlation 1] Window Help								
Linear correlation 1 Parameters Define status 1 Cross-tab parameters Sort results non Input list Cross-input (Y x X) Results Constingency Chi-Square 1 K Opfine status 2 Linear correlation 1 V X r r² t Pr(> t) Neesults Components Sort results Son Factorial analysis										
Continency Chi-Square 1 Cross-tab parameters Sort results Define status 2 More Univariate continuous stat 1 More Univariate cont stat 1 More Univariate Correlation 1 More Univariate Correlation Statistics Nonparametric statistics Instance selection Feature selection Regression Factorial analysis PLS Clustering Spv learning Spv learning assessment Scoring Association More Univariate contraction More Univariate cont stat One-way MANOVA More Univariate cont stat	Analysis					Linear corro	elation 1			
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 Bartlett's test If Group exploration Anormality Test In-Test T-Test Brown - Forsythe's test Inear correlation One-way MANOVA If Univariate continuous stat Group characterization If More Univariate cont stat Paired T-Test III Univariate discrete stat	Define status 1 Univariate discrete stat 1 Univariate discrete stat 1 Ontingency Chi-Square 1 Define status 2 With a stat		Y X NbEnfants Revenu Computation time : 0 ms. Created at 19/06/2007 13:52:20			Resul r ² 0.1847	ts t 2.0197	Pr(> t) 0.0586		
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 Association Group exploration Anormality Test T-Test T-Test Unequal Variance Fisher's test Clusterize correlation One-way MANOVA Univariate continuous stat Group characterization More Univariate cont stat Paired T-Test Univariate discrete stat				Compo	nents					
Feature selection Regression Factorial analysis PLS Clustering Spv learning Meta-spv learning Spv learning assessment Scoring Association Sartlett's test Group exploration Anormality Test T-Test T-Test Srown - Forsythe's test Clustering One-way ANOVA T-Test Unequal Variance Srown - forsythe's test Clustering More Univariate cont stat Paired T-Test	Data visualization	Statistics	Nonp	arametric statistics	Instar	nce selectio	n F	eature construction		
Spv learning Meta-spv learning Spv learning assessment Scoring Association Association Anormality Test T-Test T-Test T-Test T-Test Brown - Forsythe's test Cone-way ANOVA T-Test T-Test T-Test Fisher's test Cone-way ANOVA Cone-way ANOVA T-Test Group characterization Cone-way ANOVA Cone-way Continuous stat	Feature selection	Regression	F	actorial analysis		PLS		Clustering		
Image: Service Control Contro	Spv learning	Meta-spv learning	Spv l	earning assessment		Scoring		Association		
	ᄶ Bartlett's test Group explo 주 Brown - Forsythe's test 주 Levene's tes 주 Fisher's test 같니near corro 때 Group characterization 述More Univa		tion <u>A</u> Normality Test in One-way ANOVA ition i a One-way MANOVA te cont stat II Paired T-Test			Ian, T-Test Ian, T-Test IAN Univariate continuous stat IIII Univariate discrete stat				

If the association is monotonic but not linear, we can use nonparametric statistics like SPEARMAN or KENDALL indicators (see NONPARAMETRIC STATISTICS tab).

Association between categorical and continuous variables

We can also compute the association between categorical and continuous variables. For instance, is the OPINION different according to the number of children?

We insert the DEFINE STATUS component into the diagram. We set OPINION as TARGET, NBENFANTS as INPUT.



Then we add the GROUP CHARACTERIZATION component (STATISTICS tab).



We find that women with a definite opinion, good or bad, have fewer children on average (conditional average is respectively 2.25 and 2.0, whereas the global average is 2.7).

Conclusion

When we present the results of data mining study, basic indicators, tables and graphs, are often at least as relevant as the complicated statistical methods, fairly obscure for non-specialists. For this reason, descriptive statistics have always a large place in the reports.