1 Introduction

Statistical analysis with GNUMERIC spreadsheet.

The spreadsheet is a valuable tool for data scientist. This is what the annual KDnuggets polls reveal during these last years where Excel spreadsheet is always well placed (2017, 2016, 2015, 2014, 2013, 2012, 2011, 2010, 2009). In France, this popularity is largely confirmed by its almost systematic presence in job postings related to the data processing (statistics, data mining, data science, big data/data analytics, etc.). Excel is specifically referred, but this success must be viewed as an acknowledgment of the skills and capabilities of the spreadsheet tools.

This is not surprising. The spreadsheet is very simple to use. It has multiple features, including manipulating data tables of moderate size (e.g. 1,048,575 observations and 16,384 variables for Excel). Everyone knows to use it, at least concerning the basic features. However, computer scientists and statisticians sometimes consider it with suspicion. Some are particularly bitter (e.g. « <u>The Risks of Using Spreadsheets for Statistical Analysis</u> », IBM SPSS Statistics; as if by chance, the paper is written by an editor of statistical tool). This is somewhat simplistic. It should not be forgotten that Excel was not specifically designed to perform statistical calculations. It is not very fair to judge it exclusively in this point of view. Simply, it is important to define clearly what it can do in our context.

Precisely, Excel is widely used, but rarely separately. As indicated by the KDNuggets polls, it is operated in conjunction with specific data mining software that has the desired precision. The sharing of roles is established from this perspective: the data preparation and pre-treatment is carried out under spreadsheets; the statistical treatments are done using the specialized tools. Thus, some software vendors propose extensions (add-ins, add-ons, packages) which add additional menus and/or functions devoted to the statistical and data mining processing. <u>SAS</u> provides this kind of feature, <u>Microsoft</u> also. It is also undeniable that the use of SIPINA and TANAGRA has been largely favored by the add-ins facilitating the exchange of data with Excel and Libre/Open Office Calc.

This tutorial is devoted to the **Gnumeric Spreadsheet 1.12.12** (<u>http://www.gnumeric.org/</u>). It has interesting features: Setup and installation programs are small because it is not part of an office suite; It is fast and lightweight; It is dedicated to numerical computation and natively

¹ The French version of this tutorial was written in **May 2014**. There was a version available for Windows on that date. That is no longer the case today (the last version with Windows binaries was 1.12.17, August 2014).

incorporates a "statistics" menu with the common statistical procedures (parametric tests, non-parametric tests, regression, principal component analysis, etc.); and, it seems more accurate than some popular spreadsheets programs (McCullough, 2004; Keeling and Pavur, 2011). These last two points have caught my attention and have convinced me to study it in more detail. In the following, we make a quick overview of Gnumeric's statistical procedures. If it is possible, we compare the results with those of Tanagra 1.4.50.

We note that we use the version for Windows, but a version for Linux is also available (Figure 2). The GUI (graphical user interface) and the operating mode are the same.

2 Dataset

🐻 credit_ap	proval.xlsx - Gn	umeric	-				-			
File Edit	View Insert	Format Too	ls Statistics	Data Help						
		2 🕺 🗉	Descri Sampl	ptive Statisti ing	cs	2 3	10	0% 💌		
Arial 10	AA		Depen One S	dent Observ ample Tests	ations		= = [Až A <u>z</u>	▼ 🕅	• A •
C16	4	2	Multipl	e Sample Tests e Sample Te	sts	\$				
) A	В	С	D	E	F	G) Н		J
1	reason	guarantee	insurance	male.wage	female.wage	inc.household	family.size	inc.per.head	age	acceptation 🔺
2	Furniture	yes	yes	1238	1021	2259	2	1130	31	no
3] HiFi	yes	yes	2398	1740	4138	2	2069	43	yes
4	Furniture	no	yes	1941	1228	3169	2	1584	54	yes
5	Furniture	yes	yes	1740	1579	3319	4	830	30	yes
6	Furniture	yes	yes	1926	1426	3352	3	1117	37	yes
7] HiFi	yes	yes	1378	1653	3031	2	1516	28	yes
8	Furniture	yes	yes	2230	1316	3546	2	1773	50	yes
9] HiFi	yes	yes	2307	1674	3981	5	796	41	yes
10	Furniture	yes	yes	2236	2154	4390	4	1098	45	yes
11	Furniture	yes	yes	3492	2088	5580	2	2790	44	yes
12	Furniture	yes	no	927	1600	2527	4	632	25	no
13	Furniture	yes	yes	1566	1400	2966	4	742	35	yes
14	Furniture	yes	yes	1361	1571	2932	3	977	53	yes
15] HiFi	yes	yes	1500	896	2396	5	479	46	yes
16] HiFi	yes	yes .	2600	3107	5707	4	1427	30	yes ≡
17] HiFi	yes	yes	2600	2833	5433	4	1358	30	yes
18) HiFi	yes	no	1799	1496	3295	5	659	36	yes
19] HouseHold	yes	yes	2540	1335	3875	4	969	40	yes
20	Furniture	yes	no	1909	1178	3087	3	1029	47	no
21	Furniture	yes	yes	2976	1753	4729	4	1182	36	yes
22] HiFi	yes	yes	947	1226	2173	2	1086	56	no
23	Furniture	yes	yes	1442	734	2176	3	725	27	yes
24	HouseHold	yes	yes	834	1399	2233	4	558	35	yes
25	Furniture	yes	yes	1063	1257	2320	2	1160	36	no
26) HiFi	yes	yes	2266	1499	3765	2	1882	55	yes
27] HiFi	yes	yes	1127	1661	2788	4	697	37	no
28] HiFi	yes	yes	1425	1001	2426	3	809	26	no
29	HouseHold	yes	no	778	964	1742	2	871	65	no
30] HiFi	yes	no	459	480	939	2	470	34	no
31] HiFi	yes	yes	1229	2000	3229	4	807	43	yes 👻
	<									•
dataset								Sum =	= 0	

The "credit_approval.xlsx" data file described n = 30 loan applicants.

Figure 1 – Main window of Gnumeric, with the "Statistics" menu

We have p = 9 variables (5 quantitative, 4 categorical): reason, guarantee, insurance, male.wage, female.wage, inc.household, family.size, inc.per.head, age, acceptation (decision of the lending institution).

Book1.g	numeri	c - Gnum	eric						Fr	t, 🖂	3 4)) 16	:10 垈
Ó		<u>-</u>		8			• 7	v	🔊 Σ	f(x)	2 :12	٣
	Sans	10 🧟	a	<u>d</u> =		÷a→		C.D	% •	<u>ہ</u> ۔ ۵		*
	A1		\$) 🛛 🗸	· =							
		A) В	C	D	E	F	G) H		J	K
	1		1									
	2		T									
	3											
=	4											
	5											
	9											
	10											
	11											
	12											
	13											
	14											
Δ												
	17											
	18											
	20											
	20											
	22											
Heat	23											
		1	-						-			
	Eeu	ille1 F	euille2						Son	- 0		
	100		Contez						3011	inte = 0		

Figure 2 - Gnumeric under Ubuntu

In the following sections, we describe several Gnumeric's statistical procedures, with the same steps: how to organize the data to carry out the treatments; how to set the parameters; how start the process; and how to read the results.

3 Statistical analysis with Gnumeric

3.1 Descriptive statistics

We want to calculate various descriptive statistics for numeric variables. We copy them into the sheet named "ex.1 – desc. Stat", then we select the data range, including the header which corresponds to variable names ("Labels" in the Gnumeric terminology). We click on the Statistics / Descriptive Statistics / Descriptive Statistics menu. A dialog setting appears:

🤞 credit_app	proval.xlsx - (Gnumeric		-	-					
File Edit \	/iew Inser	t Format	Tools St	tatistics	Data Help					
🗋 🖹				Descript Samplin Depend One Sar	ive Statisti ig ent Observ nple Tests	cs ations	Correlation Covariance Descriptive Statistics			
F31	6	> 🐰	4.	Two San Multiple	nple Tests Sample Te	sts	Ranks And Percentiles			
	A	В	C	D	E	F	G H I J K L I			
	male.wage	emale.wag	c.househo	family.size	e nc.per.hea	age				
2	1238	1021	2259	2	1130	31				
3	2398	1740	4138	2	2069	43				
4	1941	1228	3169	2	1584	54	Descriptive Statistics			
5	1740	1579	3319	4	830	30				
6	1926	1426	3352	3	1117	37	Input Statistics Output			
7	1378	1653	3031	2	1516	28				
8	2230	1316	3546	2	1773	50	Input range: 'ex.1 - desc. stat.'!\$A\$1:\$F\$31 🖳			
9	2307	1674	3981	5	796	41				
	2236	2154	4390	4	1098	45	Grouped by: Columns 🗾			
	3492	2088	5580	2	2790	44				
12	927	1600	2527	4	632	25	🔍 Rows 🔨 💙			
	1566	1400	2966	4	742	35				
14	1361	1571	2932	3	977	53				
15	1500	896	2396	5	479	46	□ Labels			
	2600	3107	5707	4	1427	30				
17	2600	2000	5433	4	1350	30				
10	1799	1490	3295	5	009	30				
20	1000	1179	2075	- 4	1039	40				
20	2976	1753	4729	1	1182	36				
22	947	1226	2173	2	1086	56				
23	1442	734	2176		725	27				
24	834	1399	2233	4	558	35				
25	1063	1257	2320	2	1160	36				
26	2266	1499	3765	2	1882	55				
27	1127	1661	2788	4	697	37				
28	1425	1001	2426	3	809	26	Help Cancel OK			
29	778	964	1742	2	871	65				
30	459	480	939	2	470	34				
31	1229	2000	3229	4	807	43				
32										
dataset ex	iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii									

Into the INPUT tab, the data range must be selected. The variables are organized by columns in our case. We select the LABELS option to specify that the first row of the data range corresponds to the variable names.



We do not change anything in the STATISTICS tab. In the OUTPUT tab, we specify the coordinates for the output. We observe the "Enter into cells: Formulae" option. It means that the results will be inserted as formulas. Thus, if the values into the data range are changed, the results will be automatically updated. This property is particularly interesting. However, Gnumeric does not automatically adapt to a change in the size of the data (additional rows and columns).

We obtain, among others, the mean, the median, the standard deviation, etc. (the results are formatted to make easy the reading).

诸 *credit_ap	🔁 *credit_approval.xlsx - Gnumeric												
File Edit \	√iew Insei	rt Format	Tools S	tatistics D)ata Help								
			(6 (3 9	- 🥐	- 8	$\sum f(x) = \frac{1}{2} \frac{z}{2}$, 👔 100)% 🔻				
Condoro	11					- 653	o7 \$0.00			m _ 4	- 4	-	
Candara		AA					70 • °.00 =.0			⊞ ▼ №		•	
13	<u> </u>	💫 🐰	<₽ 👻	= =sq	rt(var('ex.:	1 - desc. s	tat.'!\$A\$2:\$A\$31)/	count('ex.1 -	desc.stat.'!\$	A\$2:\$A\$31))			
	A	В	С	D	E	F	6 н	I	J	К	L	M	N
1	male.wage	emale.wag	c.househo	family.size	nc.per.head	age		male.wage	female.wage	inc.householc	family.size	inc.per.head	age 🔺
2	1238	1021	2259	2	1130	31	Mean	1741.1333	1508.9667	3250.1000	3.2000	1107.4022	39.8333
3	2398	1740	4138	2	2069	43	Standard Error	130.9361	100.3394	206.2781	0.1942	94.6763	1.8514
4	1941	1228	3169	2	1584	54	Median	1	1461	3128	3	1003.17	37
5	1740	1579	3319	4	830	30	Mode	200	#N/A	#N/A	2	#N/A	30
6	1926	1426	3352	3	1117	37	Standard Deviati	717.1668	549.5817	1129.8318	1.0635	518.5637	10.1407
7	1378	1653	3031	2	1516	28	Sample Variance	514328.1885	302040.0333	1276519.8862	1.1310	268908.2783	102.8333
8	2230	1316	3546	2	1773	50	Kurtosis	-0.2380	2.1551	0.2216	-1.3870	2.6041	-0.2180
9	2307	1674	3981	5	796	41	Skewness	0.4127	1.0494	0.5449	0.1253	1.4493	0.5969
10	2236	2154	4390	4	1098	45	Range	3033	2627	4768	3	2320.5	40
11	3492	2088	5580	2	2790	44	Minimum	459	480	939	2	469.5	25
12	927	1600	2527	4	632	25	Maximum	3492	3107	5707	5	2790	65
13	1566	1400	2966	4	742	35	Sum	52234	45269	97503	96	33222.067	1195
14	1361	1571	2932	3	977	53	Count	30	30	30	30	30	30 =
15	1500	896	2396	5	479	46							
16	2600	3107	5707	4	1427	30							
17	2600	2833	5433	4	1358	30							
18	1799	1496	3295	5	659	36							
<u> </u>	2540	1335	3875	4	969	40							
20	1909	1178	3087	3	1029	47							
21	2976	1753	4729	4	1182	36							
22	947	1226	2173	2	1086	56							
23	1442	/34	2176	3	725	27							
24	834	1399	2233	4	558	35							
25	1063	1257	2320	2	1160	- 36 - 55							
26	2266	1499	3765	2	1882	55							
20	1/05	1001	2/00	4	697	3/ ne							
29	1420	1001	2420	<u>ა</u>	009 971	20							
30	459	480	939	2	470	34							
	4.00	400	555	~	410							·	
dataset er	v 1 • des	c stat									C	100.0001	
Lacaser es	ar - des	c. stat.									sum =	120.9301	

Let us see the standard deviation for the variable "X : male.wage". Into **I3** cell, we see the formulae $s_{\bar{x}} = \sqrt{\frac{s_x^2}{n}} = \sqrt{\frac{514328.1885}{30}} = 130.9361$. The estimated variance s_x^2 of X is into **I7**.

By comparison, we obtain the following results for "male wage" under **Tanagra 1.4.50**. The results are consistent.

TANAGRA 1.4.50 - [More	Univariate cont stat 1]						x
Tile <u>D</u> iagram Compo	nent <u>W</u> indow <u>H</u> elp					- 6	×
🗅 📽 🖬 👫							
Default ti	tle		Res	aults			
Dataset (credit_app	oroval_dataset.xls)	Attribute		Sta	ats		Ш
More Univari	ate contistat 1		Statistics				
			Average		17	41.1333	
			Median		16	53.0000	
			Std dev. [Coef of var	iation]	717.1668	[0.4119]	
			MAD [MAD/STDDEV]		589.6089	[0.8221]	
			Min * Max [Full range]		459.00 * 3492.00 [3033.00]	
		male.wage	1st * 3rd quartile [Ra	inge]	1229.00 * 2266.00 [1037.00]	
		materinage	Skewness (std-dev)		0.4127	(0.4269)	
			Kurtosis (std-dev)		-0.2380	(0.8327)	÷
		•	III			Þ	
) [_	Componer	nts				
Data visualization	Statistics	Nonpara	metric statistics	Insta	nce selection		
Feature construction	Feature selection	Re	egression	Fac	torial analysis		
PLS	Clustering	Sp	v learning	Met	a-spy learning		
Spv learning assessment	Scoring	As	sociation			1	
ANOVA Randomized Blog	ks 🔤 Box's M Test		Tisher's test		Gro	up explo	rati
Bartlett's test	Brown - Fors	ythe's test	Group chara	cterizat	ion 🫃 Hot	elling's T	2
<		-				-	Þ
							, d

3.2 Correlation tool

We use the same numeric variables to calculate the matrix of pairwise correlations coefficients. We duplicate the data range into a new worksheet "ex.2 – corr matrix". After we select the data range, we click on the **Statistics / Descriptive Statistics / Correlation** menu.

Correlation	Correlation
Input t	Input Output
Input range: 'ex.2 - corr matrix'!\$A\$1:\$F\$31 🖳 Grouped by: Columns Rows	Output Placement New sheet New workbook Output range: 'ax 2 - corr matrix!!!!
© Areas ☑ Labels	Output Formatting Autofit columns Clear output range Retain output range formatting
Help Cancel OK	Enter into cells: Formulæ

We check the input range and set the right parameters (variables grouped in columns, labels). Into the OUTPUT tab, we set the output range.

🤠 credit_ap	g credit_approval.xlsx - Gnumeric												
<u>E</u> ile <u>E</u> dit	<u>F</u> ile <u>E</u> dit <u>V</u> iew Insert F <u>o</u> rmat <u>T</u> ools <u>S</u> tatistics <u>D</u> ata <u>H</u> elp												
	🗅 🖻 🗔 📇 🤮 🕌 🖻 📋 🔶 🔻 🦿 - 🌍 🗵 f(0) 🦕 🐝 🌆 100% 💌												
Candara	Candara 10 🔺 🔺 📄 📰 📰 📰 📰 🐯 % 🔸 🐝 🔐 📻 📰 👫 🥂 🗸 🗸												
J4	4 💫 💥 🖑 🔻 = ecorrel('ex.2 - corr matrix'!\$B\$2:\$B\$31,'ex.2 - corr matrix'!\$C\$2:\$C\$31)												
	Correlations	male.wage	female.wage	inc.householc	family.size	inc.per.head	age						
2	male.wage	1											
3	female.wage	0.5837	1										
4	inc.household	0.9187	0.8570	1						- =			
5	family.size	0.1334	0.3	0.2404	1								
6	inc.per.head	0.6760	0.394	0.6211	-0.5507	1							
	age	0.0390	-0.1595	-0.0528	-0.3293	0.2637	1	1					
9													
12													
13													
14													
15										-			
	₹							1		•			
● e:	ex.2 - corr matrix Sum = 0.8570												

We note that the correlation coefficients are obtained with the **CORREL** function.

3.3 Principal component analysis (PCA)

We create a third sheet "ex.3 – PCA" and we copy the numeric variables. We select the data range, then we click on the **Statistics / Dependent Observations / Principal Components Analysis** menu. We set the following parameters:

Principal Components Analysis	🐻 Principal Components Analysis
Input	Input Output
Input range: ex.3 - PCA'!\$A\$1:\$F\$31 Grouped by: Columns Rows Areas	Output Placement New sheet New workbook Output range: 'ex.3 - PCA'!H1 Output Formatting Autofit columns Clear output range Retain output range formatting
Help <u>Cancel</u>	Retain output range comments Enter into cells: Formulæ Help Cancel OK

We obtain the following results:

🤠 credit_ap	proval.xlsx - Gr	umeric	-	_	_			3	
File Edit \	View Insert	Format To	ols Statisti	cs Data He	alp				
				<u> </u>	_ 0		a, z,	_	
		<u>io</u> //	••••••••	2 - C	· · · · 🕹	<u>≥</u> ƒ(<u>x</u>)	≌`z ≌`a	•	
Candara	10 🔺	AA		7	88 🐶	% · +	0.00	•	
114	<u>ک</u>		• =	{=eigen(I10)/(110-1)*13:1	18)}(6,7)[0]	0]		
								_	
	ЦН		J	K j	L	M		£	
	Principal Com	ponents Anal	ysis						
2	Covariances	male.wage	female.wage	nc.householc	family.size	inc.per.head	age	-	
3	male.wage	497183.92	222406.64	719590.55	98.34	243023.85	273.99		
4	female.wage	222406.64	291972.03	514378.67	180.91	108756.98	-859.07		
5	inc.household	719590.55	514378.67	1233969.22	279.25	351780.83	-585.08		
6	family.size	98.34	180.91	279.25	1.09	-293.59	-3.43		
7	inc.per.head	243023.85	108756.98	351780.83	-293.59	259944.67	1340.43		
8	age	273.99	-859.07	-585.08	-3.43	1340.43	99.41		
9									
10	Count	30	30	30	30	30	30		
11	Mean	1741.13	1508.97	3250.10	3.20	1107.40	39.83		
12	Variance	514328.19	302040.03	1276519.89	1.13	268908.28	102.83		
13									
14	Eigenvalues	2049615.200	205859.781	106338.397	86.819	0.154	0.000		
15	Eigenvector	0.46922	-0.39587	-0.53832	-0.00080	-0.00047	0.57735	Ξ	
16		0.31710	0.59935	0.45484	0.00295	-0.00013	0.57735		
17		0.78631	0.20348	-0.08347	0.00215	-0.00059	-0.57735		
18		0.00013	0.00162	-0.00195	0.00160	1.00000	0.00000		
19		0.24697	-0.66527	0.70451	-0.00855	0.00243	0.00000		
20		-0.00014	-0.00822	0.00443	0.99996	-0.00157	0.00000		
21									
22		ξ	ξ2	5	ξ4	ξ5	ξ6		
23	male.wage	0.93668	-0.25045	-0.24477	-0.00001	0.00000	0.00000		
24	female.wage	0.82603	0.49481	0.26988	0.00005	0.00000	0.00000		
25	inc.household	0.99636	0.08171	-0.02409	0.00002	0.00000	0.00000		
26	family.size	0.17025	0.68961	-0.59902	0.01398	0.36937	0.00000		
27	inc.per.head	0.68184	-0.58208	0.44303	-0.00015	0.00000	0.00000		
28	age	-0.01944	-0.36761	0.14246	0.91880	-0.00006	0.00000		
29									
30	Percent of Tr	86.78%	8.72%	4.50%	0.00%	0.00%	0.00%	-	
· · · ·	•						4		
◀ ex.	ex.3 - PCA Sum = 2049615.200								

We read successively:

- The covariance matrix. Gnumeric uses the covariance matrix for PCA.
- The number of observations per variables, the means and the variances.
- The eigenvalue for each factor (component).
- The eigenvectors.
- The correlations of the variable with the components.
- The proportion of the variance represented by each factor.

There is not an option to perform a PCA based on the correlation matrix. A simple solution is to replace the values of the covariance by the correlation i.e. instead the COVAR function, we use the CORREL function in **I3:N8**. The subsequent results are updated automatically. I find this possibility quite exciting.

By comparing the results with those of Tanagra, I notice a slight difference about the eigenvalues (below the Tanagra's results):

Axic		Difforonco	Proportion	Cumulative
AXIS	Eigen value	Difference	(%)	(%)
1	1981294.694	1782296.905	86.78%	86.78%
2	198997.789	96204.006	8.72%	95.49%
3	102793.782	102709.858	4.50%	100.00%
4	83.925	83.776	0.00%	100.00%
5	0.149	0.149	0.00%	100.00%
6	0	-	0.00%	100.00%
Tot.	2283170.339	-	-	-

We find the explanation of this difference in the formula used by Gnumeric (the cell and the formulae are highlighted by arrows into the screenshot above). Gnumeric displays $\frac{n}{n-1} \times \lambda_1 = \frac{30}{29} \times 1981294.694 = 2049615.2$ where n = 30 is the number of instance, λ_1 is the first eigenvalue of the covariance matrix. The eigenvectors are weighted in the same way. The correction becomes negligible when n is large ($\frac{n}{n-1} \rightarrow 1$). But, nevertheless, the correlations between the variables and the factors are not modified. This is the most important thing when we want to interpret the results.

3.4 Linear regression

We want to explain the family size from the income and the age (I know that the example is a bit crazy, the aim of the tutorial is to show how to use Gnumeric and not to perform a relevant analysis). We copy the dataset into a new sheet "**ex.4 – Linear reg**". We put in the order the variables: "inc.household", "age" and "family.size".

We click on the **Statistics / Dependent Observations / Regression** menu. We set the following parameters:

Regression	Regression	Regression
Input Output	Input Options	Input Options Output
 Multiple linear regression 	Confidence level: 0.95	Output Placement
Multiple 2-variable regressions	Force intercept to be zero	New sheet
Multiple dependent (y) variables		New workbook
Vyoriahlasy Jay 4, Linear regultAdt1,tDt21	Calculate residuals	● Output range: 'ex.4 - Linear reg'!E1 🖳
X Variables: Tex.4 - Linear Teg?(\$A\$1:\$B\$31 Hz		Output Formatting
Y variable: 'ex.4 - Linear reg'!\$C\$1:\$C\$31 🖳		Autofit columns
☑ Labels		Clear output range
1		🔲 Retain output range formatting
		🔲 Retain output range comments
		Enter into cells: Formulæ
Help Cancel OK	Help Cancel OK	Help Cancel OK

Y is the dependent variable (familiy.size), X is the range of the explanatory variables (inc.household and age). Gnumeric uses the LINEST function. It reorganizes the results for a presentation consistent with the standard statistical tools. It picks the different values in an internal table with the INDEX function.

诸 *credit_a	pproval.xlsx - Gr	numeric	-	-	_		-				-		X
File Edit	∨iew Insert	Format Too	ols Statistics	Data Help									
			ē 🖻 🔤	▼ 🤶	- 🕹 🛛	∑ f(x) ⁿ ⊻z	z Yn 🚺	100%	•				
Candara	10 🖪 🖊				3 🐶 🕺	• \$00	.00	*= A	A <u>x</u>	⊞ ▼	<mark>3n</mark> ▼	<u>A</u> •	
F17	4	💥 🖑	• = {=	inde×(linest	('ex.4 - Line	ar reg'!\$C\$2	:\$C\$31,'ex.4	l - Linear I	reg'!\$A\$2	2:\$B\$31,	TRUE, TRUE	;),1,3)}(1,1)[0][0]
) E	F	G	н	I	J	K	L	M	[N	0) P	
1	SUMMARY OUT	TPUT	Response Varia	family.size		Î							A
2													
3	Regression Sta	itistics											
4	Multiple R	0.397913											
5	R^2	0.158335											
6	Standard Erro	1.011171											_
7	Adjusted R^2	0.095990											E
8	Observations	30											_
9]									_			
10	ANOVA									_			
		df	55	MS	F	gnificance of	F						
12	Regression	2	5.1934	2.5967	2.5396	0.0976							
13	Residual	27	27.6066	1.0225									
14	Total	29	32.8										
15													
16		Coefficients	Standard Error	t-Statistics	p-Value	Lower 95%	Upper 95%						
	Intercept	3.842254	0.956243	4.018074	0.000422	1.880206	5.804303						
18	inc.household	0.000211	0.000166	1.264897	0.216717	-0.000131	0.000552			_			
19	lage	-0.033300	0.018542	-1.795865	0.083716	-0.071345	0.004746						
20	4												
I ex.	.3 - PCA	ex.4 -	Linear reg							Sum = 3	3.842254		

We observe: the overall results (coefficient of determination R², standard error of residuals, etc.); the analysis of variance (ANOVA) table, the F-Test for global significance of the regression; and finally, the table of coefficients, with their standard error, the t statistic for the testing that the population regression coefficient for each variable is equal to zero, the observed significance level (p-value), the confidence interval for the coefficient at 95% confidence level.

We obtain the same results with Tanagra. The organization is identical.

TANAGRA 1.4.50 - [Multiple linear regression 1]		1.1	1.4.1	4.1			ĸ		
Tile Diagram Component Window Help						- 8	×		
D 📽 🔲 👫									
Default title Report (X'X)^(-1) matrix									
Global results									
More Univariate cont stat 1	Endogenous att	ribute		family.size					
Principal Component Analysis 1	Examples			30			_		
🖻 📲 Define status 2	R ²			0.158335					
Multiple linear regression 1	Adjusted-R ²			0.095990					
1	Sigma error			1.011171					
	F-Test (2,27)		2.539	6 (0.097584)			=		
	Analysis o	of variand	e						
	Source	xSS	d.f.	xMS	F	p-value			
	Regression	5.1934	2	2.5967	2.5396	0.0976			
	Residual	27.6066	27	1.0225					
	Total	32.8000	29						
	Coefficie	nts							
	Attrit	oute	Coef.	std	t(27)	p-value			
	Intercept		3.842254	0.956243	4.018074	0.000422			
	inc.household		0.000211	0.000166	1.264897	0.216717			
	age		-0.033300	0.018542	-1.795865	0.083716	-		
	•	111				۰.			
	Co	omponents							
Data visualization Statistics	Nonparam	etric statistics	Instance	e selection	Feature of	construction			
Feature selection Regression	Factor	ial analysis	1	PLS	Clu	stering			
Spv learning Meta-spv learning	Spv learning assessment Scor			oring	Asso	ciation	Ì		
🛃 Backward Elimination Reg 🛛 DfBetas	🖳 Forv	ward Entry Reg	ression 🛛 🔀 Nu	SVR	°≟i R	egression Asses	sm		
🚓 C-RT Regression tree 🛛 🥳 Espilon SVR	🛃 Mul	tiple linear reg	ression 🛛 🖄 Ou	tlier Detection	елан Каралан Ка Каралан Каралан К	egression tree			
·							۴		

3.5 One-sample t-test

We wonder whether the man and the woman in the same household have comparable wages. For this purpose, we copy the two columns in a new sheet "**ex.5** – **One sample t-test**". We create a new variable DIF which is calculated from the difference (male.wage - female.wage). Under the null hypothesis, the wages are identical, this difference should be equal to 0 on average. Thus, we perform a comparison to a nominal mean which is 0.

After we select the column DIF, we click on the **Statistics / One Sample Tests / Claims About a Mean** menu. We set the following parameters into the dialog setting.

Tanagra

Ricco Rakotomalala

Claims About a Mean	Claims About a Mean	Claims About a Mean
Input Output	Input Test put	Input Test Output
Input range: Dne sample t-test!!\$C\$1:\$C\$31	Predicted Mean: 0	Output Placement
Grouped by: Columns	Alpha: 0.05 🗖 🕂	New sheet
Bows		New workbook
Areas		Output range: - One sample t-test'!E1
		Output Formatting
		🗹 Autofit columns
		🗹 Clear output range
		Retain output range formatting
		Retain output range comments
		Enter into cells: Formulæ
Help Cancel OK	Help Cancel OK	Help Cancel OK

At the 5%, we reject the null hypothesis². We see at the cell **F6** the formulae ($\mu_0 = 0$ for our example, this the Predicted Mean option into the TEST tab):

$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}$$

The p-value is obtained with the TDIST function.

🍓 *credit_ap	诸 *credit_approval.xlsx - Gnumeric												
File Edit View Insert Format Tools Statistics Data Help													
	🗋 🛅 🔚 🗸 🚜 🗊 🔯 🥠 🔻 🦿 🔪 🕱 fox 😓 🐝 💌												
Candara 12 🔺 🔺 🖨 📄 🗐 🗐 🖽 💷 🌄 🐯 % • 🀝 🔐 🛛 🗸													
F6 🔷 💥 🛫 = {=(F3-F4)/sqrt(F5/F2)}(1,1)[0][0]													
	A	В	C	D	E	[F]							
1	male.wage	female.wage	dif		Student-t Test	dif							
2	1238	1021	217		N	30							
3	2398	1740	658		Observed Mean	232.1667	-						
4	1941	1228	713		Hypothesized Mean	0							
5	1740	1579	161		Observed Variance	356216.5575							
6	1926	1426	500		Test Statistic	2.1306							
7	1378	1653	-275		df	29							
8	2230	1316	914		α	0.05							
9	2307	1674	633		P(T≤t) one-tailed	0.0209							
10	2236	2154	82	- (P(T≤t) two-tailed	0.0417							
11	3492	2088	1404										
12	927	1600	-673										
10	4500	4.400	4.00				1						
ex.5 • One	e sample t-	test			Sum = 2.13	806	ex.5 - One sample t-test						

² Another approach is to calculate the ratio of the male and female wages, and the compare the mean of the new variable with 1. We obtain the same conclusion but the test scheme is different [ex.5 (bis) – One sample t-test].

3.6 Comparison of two means – Paired samples

Another way to compare the male and female wages inside the household is to conduct a comparison of means between paired samples³. We copy the two columns of wages into the new sheet "ex.6 – Paired t-test". We click on the Statistics / Two Sample Tests / Claims About Two Means / Paired Samples menu.



We select the data range into the INPUT tab, with the **Labels** option. The two columns of data must have the same length, otherwise the comparison is not possible.

³ https://onlinecourses.science.psu.edu/stat500/node/51

🤠 credit_app	🔁 credit_approval.xlsx - Gnumeric								
File Edit 🕚	File Edit View Insert Format Tools Statistics Data Help								
	🗋 📄 🗔 📇 🖧 🦂 🛱 🔯 🤌 🔻 🦿 👻 🏹 🚳 🗵 f(x) 💱 🐝 🗎 🏛 100% 🔽								
Candara 12 🗛 🗛 📋 🗉 🗉 🖽 🖽 🍞 🐯 % • 號 🔐 📻 🖅 👻									
E10	E10 💫 💥 🛫 = =(E7-E6)/(E8/(E9+1))^0.5								
	A	В	C	D	E	F			
1	male.wage	female.wage			male.wage	female.wage			
2	1238	1021		Mean	1741.1333	1508.9667			
3	2398	1740		Variance	514328.1885	302040.0333			
4	1941	1228		Observations	30	30	Ξ		
5	1740	1579		Pearson Correlation	0.5837				
6	1926	1426		Hypothesized Mean Difference	e o				
7	1378	1653		Observed Mean Difference	232.1667				
8	2230	1316		Variance of the Differences	356216.5575				
9	2307	1674		df	29				
10	2236	2154		t Stat	2.1306				
11	3492	2088		P (T<=t) one-tail	0.0209				
12	927	1600		t Critical one-tail	1.6991				
13	1566	1400		P (T<=t) two-tail	0.0417				
14	1361	1571		t Critical two-tail	2.0452		-		
· - · ·	•						•		
	x.6 - Paire	d t-test			Sum = 2.130	6	//		

By a different process, we obtain exactly (the values and distributions of the test statistics are identical) the same result as previously (section **Erreur ! Source du renvoi introuvable.**). Men and women inside a household have not the same wages in average. It seems that men are advantaged.

3.7 Non-parametric test – Paired samples

We can answer to this question - is the wages of the men and the women are equivalent inside the household - by the means of a non-parametric test. We use Wilcoxon signed rank test. The test statistic is based on the rank of the differences. We do not need that the populations are assumed to be normally distributed.

Ricco Rakotomalala

Tanagra

Claims About Two Medians (Paired Samples)	Claims About Two Medians (Paired Samples)	Claims About Two Medians (Paired Samples)
Input Output	Input Test	Input Test Output
Variable 1 range: 'ex.7 - Paired Wilcoxon'!\$A\$1:\$A\$31 Variable 2 range: 'ex.7 - Paired Wilcoxon'!\$B\$1:\$B\$31 Variable 2 range: 'ex.7 - Paired Wilcoxon'!\$B\$1:\$B\$31 Labels	 Sign Test Wilcoxon Signed Rank Test Hypothesized difference of medians: 0 Alpha: 0.05 	Output Placement New sheet New workbook Output range: Utput rormatting Autofit columns Clear output range Retain output range formatting Retain output range comments Enter into cells:
Help Cancel OK	Help Cancel OK	Help Cancel OK

We copy the two columns in a new sheet "ex.7 – Paired Wilcoxon". We click on the Statistics / Two Sample Tests / Claims About Two Medians / Wilcoxon Signed Rank Test menu. After we set the needed settings (see screenshot above), we obtain:

🤠 credit_app	👸 credit_approval.xlsx - Gnumeric									
File Edit View Insert Format Tools Statistics Data Help										
$ \begin{tabular}{ c c c c c } \hline \begin{tabular}{ c c c c c } \hline \begin{tabular}{ c c c c } \hline \begin{tabular}{ c c c c c c c } \hline \begin{tabular}{ c c c c c c c c c c c c c c c c c c c$										
Candara 12 🔺 🔺 📋 📳 🗄 🖽 🖽 🛤 🐺 🗸 · 號 🔐 🐺 😁 🖓 🗛 🎛 🔻 🗸										
E10	4	> 🗶 🖑	▼ = =if(E5<	12,#N/A,norm	ndist(E8+0.5,E5*(8	E5+1)/4,sqrt(E	5*(E5+1)/	4*(2*E5+	1)/6),TRUE	=))
	A	В	C) D		E	F	G	Н		
1	male.wage	female.wage	Wilcoxon Signe	d Rank Test	male.wage	female.wage				-
2	1238	1021	Median		1653	1461				
3	2398	1740	Observed Medi	an Difference	191.5					=
4	1941	1228	Predicted Medi	an Difference	0					_
5	1740	1579	N		30					
6	1926	1426	S-		144					
7	1378	1653	S+		321					
8	2230	1316	Test Statistic		144					
9	2307	1674	α		0.05	\sim				
10	2236	2154	P(T≤t) one-taile	d	0.035147					
11	3492	2088	P(T≤t) two-taile	P(T≤t) two-tailed						
12	927	1600						-		
(¹	▲ FCC	4.400	1		III			1		•
◀ ex.6 - F	Paired t-tes	t ex.7 - P	aired Wilcoxon 🕽			Sur	n = 0.035	147		1

The sample size being enough ($n \ge 12$), Gnumeric provides the p-value based on the normal distribution by calculating the Z value. For a two-sided test, we have p-value = 0.070294. The differences between the wages is less obvious with this procedure.

In comparison of the outputs of Tanagra, we observe a slight difference (Z = 1.820298, p-value =0.068714).

Attribute_Y		Attrib	ute_X	Statistical test		
male.wage		female	e.wage	Measure	Value	
Avg	1741.133333	Avg	1508.966667	Used examples	30	
Std-dev	717.166779	Std-dev	549.581689	Sum ranks + (T+)	321	
				Sum ranks - (T-)	144	
				E(T+)	232.5	
				V(T+)	2363.75	
				Z	1.820298	
				Pr(> Z)	0.068714	

The divergence relies on the Gnumeric's use of the continuity correction. It calculates Z'

$$Z' = \frac{|T^+ - E(T^+)| - 0.5}{V(T^+)} = \frac{|312 - 232.5| - 0.5}{\sqrt{2363.75}} = 1.810014$$

Using the cumulative distribution function $\Phi(.)$ of the standardized normal distribution,

 $p.value = 2 \times [1 - \Phi(1.810014)] = 0.070294$

This is the p-value provided by Gnumeric.

3.8 Comparison of means – Independent samples

Our aim is to compare the "income per head" of the household according to the "acceptation" i.e. the decision of the lending institution. We have a partition of the dataset in two independent samples. This kind of statistical test needs a specific formatting under Gnumeric. We create a new sheet "ex.8 – indep parametric". Instead of the usual organization of the data, we need to create two columns for the values of "inc.per.head" according to the values of "acceptation", one for each modality (yes or no). These two columns have not necessarily the same length.

🐻 credit_ap	📴 credit_approval.xlsx - Gnumeric							
File Edit View Insert Format Tools Statistics Data Help								
Inc.per.head for							•	
Acce	ptatic	on = n	0	Inc	.per.h	nead f	for	•
A12		i 🔊 🖓	🥮 🔻 👘	Acc	ontat	tion -	NOC	
		В		ACC	epta	uon –	yes	
2	1 470	179 J						t—ĥ
3	632	558						
4	697	659						<u> </u>
5	809	725						
6	871	742						
7	1029	796						
8	1086	807						
9	1130	830						E
10	1160	969						
11]	977						
12		1098						
13)	1117						
14]	1182						
15]	1358						
16	ļ	1427						
17	ļ	1516						
18	ļ	1584						
19	Į	1773						
20	Į	1882						
21	Į	2069						
22	Į	2790						
23	4							
indep par	ametric				Sum :	= 0		

We see below the settings for the procedure **Statistics / Two Sample Tests / Claims about two means / Unpaired Samples, Unequal Variances**. We assume that the two populations have unequal variances.

Claims About Two Means		Claims About Two Means		
Input Input Input		Input Populations Output		
Variable 1 range: k.8 - indep parametric!!\$As	\$1:\$A\$10 🖳	∨ariables are:	🔘 Paired	Onpaired
Variable 2 range: 0.8 - indep parametric!!\$B	\$1:\$B\$22	Population variances are:	🔘 Known	Onknown
		Population variances are:	🔘 Equal	Onequal
Help	ок	Help	Cancel	ОК
Claims About Two Means		Claims About Two Means		
Input Populations Test		Input Populations Test Output		
Hypothesized mean difference: b		Output Placement		
Alpha: 0.05		New sneet		
		Output range: 'ex.8 - inde	ep parametrio	
		Output Formatting		
		Clear output range		
		Clear output range Rotain output range	ttipa	
		Retain output range comm	unig ants	
		Enter into cells: Eormulas	ienco	
Help	ок	Help	<u>C</u> ancel	<u>о</u> к

The procedure is based on the Welch's t-test (<u>http://en.wikipedia.org/wiki/Welch's_t_test</u>). The test statistic is easy to calculate. The main issue is to calculate properly the degrees of freedom of the Student distribution in this context.

Gnumeric provides the following results. The degree of freedom is not an integer value (d.f. = 27.99). The observed p-value is equal to 0.035396 for a two-sided test.

😼 *credit_approval.xlsx - Gnumeric							
File Edit View Insert Format Tools Statistics Data Help							
🗅 🖻 🖥 📇 🗸 🖡 🗊 😰 🥠 🔨 💌 🤤 🗵 f(x) 😓 💑 🗎 100% 🔽							
Candara 12 🗛 🛆 📄 🗐 📰 🎦 🏦 😽 % • 🐝 🔐 📻 😁 📔 🔻							
E7 $\implies \qquad $							
	A	В	C	D)[E]	F]	
1	no	yes			no	yes	-
2	470	479		Mean	875.87963	1206.62619	
3	632	558		Variance	59267.18248	331751.24499	
4	697	659		Observations	9	21	
5	809	725		Hypothesized Mean Differe	e o		
6	871	742		Observed Mean Difference	-330.74656		
7	1029	796		df	27.99		
8	1086	807		t Stat	-2.210736		Ξ
9	1130	830		P (T<=t) one-tail	0.01770		
10	1160	969		t Critical one-tail	1.70115		
11		977		P (T<=t) two-tail	0.035396		
12		1098		t Critical two-tail	2.04844		
13		1117					
14		1182					
15		1358					
16		1427					
1/		1516					
18		1584					
20		1773					
20		2069					
22		2790					
	4						- T
◀ ex.8	- indep	parameti	ric 🕨		Sum = 27.99		

Compared with Tanagra, we have the same results except for the p-value.

Attribute_Y	Attribute_X		Descr		Statistical test		
		Value	Examples	Average	Std dov		-330.7466 / 149.6092 =
	value Examples		Average	Sta-dev	т	-2.210736	
inc.per.head	acceptation	no	9	875.8796	243.4485	d.f.	27.99
		yes	21	1206.6262	575.9785	p-value	0.035393
		All	30	1107.4022	518.5637		

This difference on the p-value is explained by the treatment of the degrees of freedom. Tanagra truncates the value by dropping all decimal places (df = 28 for this example). The TDIST function of Gnumeric seems to use a linear interpolation between the nearest integer values (df between 27 and 28 for our example). But the gap between the p-value is very low anyway.

3.9 Non-parametric test – Independent samples

We use the Wilcoxon-Mann-Whitney test in this section. The dataset (sheet "ex. 9 indep non parametric") must be organized such as previously (section Erreur ! Source du renvoi

introuvable.). We click on the Statistics / Two Sample Tests / Claims About Two Medians /

Wilcoxon-Mann-Whitney test menu. We set the following settings.

🐻 Wilcoxon-Mann-Whitney Test	😼 Wilcoxon-Mann-Whitney Test
Input but	
Variable 1 range: parametric!!\$A\$1:\$A\$10	Output Placement
Variable 2 range: parametric!!\$B\$1:\$B\$22	 New sneet New workbook
✓ Labels	Output range: dep non parametric'!D1
	Output Formatting
	🖉 Autofit columns
	🔽 Clear output range
	🔲 Retain output range formatting
	🗐 Retain output range comments
	Enter into cells: Formulæ
Help Cancel OK	Help Cancel OK

The Z value used for the normal approximation is not displayed explicitly. But it is used for the calculation of the p-value with the Gnumeric's NORMDIST function.

🐻 *credit_ap	proval.xlsx -	Gnumeric		for each of	-		an 1	-			x
File Edit \	View Insei	t Format	Tools S	statistics D	ata Hel	р					
📄 🖻 🖥 📇 🕌 🛱 🖆 🔯 🤸 🐨 🦿 😒 🛣 🕬 🦕 🔻											
Candara 10 🗛 🗛 📄 🗐 🗐 🎟 🔫 🎫 😽 🐯 🕉 🕂 🏶 🔐 📮 😁 🔻											•
E9 $\implies \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ $											
	A	B	С	<u> </u>	E	F)	G[Н			
1	no	yes		Wilcoxon-M	ann-Whitn	ey Test					
2	470	479			no	yes	Total				
<u></u>	632	558		Rank-Sum	109	356	465				_
	697	659		N	9	21	30				
5	809	725		U	64	125	189				_
6	871	742		Ties	0						_
7	1029	796		Statistic	109						_
8	1086	807		U-Statistic	64						_
9	1130	830		p-Value	0.16749						- 11
10	1160	969									_
	ļ	977									
12	ļ	1098									
13	ļ	1117									_
14	ļ	1182									_
	Į	1358									_
16	ļ	1427									_
	ļ	1516									
18	ļ	1584									
	ļ	1773									
	ļ	1882									_
21	1	2069									_
22	1	2790									-
	∢										F
ex.9 - in	dep non	paramet	ric 🕨				Sum =	0.16749	9		

	Value	Examples	Average	Rank sum	Rank mean	Mann-Whitney U	64
	no	9	875.8796	109	12.1111	E(U)	94.5
inc.per.headacceptatio	n yes	21	1206.6262	356	16.9524	V(U)	488.25
	All	30	1107.4022	465	15.5	Z	1.38032
		•				P(> Z)	0.16749

Tanagra provides the same results, but it presents them differently.

3.10 One-way Analysis of variance (One-way ANOVA)

We want to compare the age of the persons according to the "reason" of the loan. We create a new sheet "**ex.10 – Anova**". We have a list of 3 columns because "reason" has 3 distinct possible values {furniture, hifi, household}. The number of observations can be different into the columns. Here is the organization of the dataset into the new worksheet.

🐻 credit_ap	proval.xlsx - G	numeric	- 1	-		• X
File Edit \	View Insert	Format	Tools Stat	istics Da	ata Help)
			E 💼		-	• •
Value	s of age	accord	lina the I	evelso	of reas	son 👻
L			<u> </u>			
A1		L		= Furnit	ure	
		В	С	D) E	F
1	Furniture	HiFi	HouseHold			
2	25	26	35			
3] 27	28	40			
4) 30	30	65			
5] 31	30				
6] 35	34				
7] 36	36				_
8] 36	37				
9	37	41				-
10	44	43				
	45	43				
12	47	46				_
13	<u> </u>	55				_
14	53	56				
15	54					
	Į					v
	<		111			
va			Sum :	= 0		//

We click on the **Statistics / Multiple Sample Tests / ANOVA / One Factor** menu to launch the analysis. Into the dialog setting, we select a rectangular data range. This is not matter if some cells are empty.

Tanagra

Ricco Rakotomalala

🐻 ANOVA -	Single Factor	ANOVA - Single Factor	🐻 ANOVA - Single Factor
Input	ons Output	Input Options	Input Options Output
Input ra	ange: 'ex.10 - anova'!\$A\$1:\$C\$15	Alpha: 0.05 🚍 🛟	Output Placement
Groups	d by: Columns		New sheet
Groupe			🔘 New workbook
	© Rows		Output range: 'ex.10 - anova'!E1
	Areas		Output Formatting
🔽 Lab	els		Autofit columns
			Clear output range
			🔲 Retain output range formatting
			📄 Retain output range comments
			Enter into cells: Formulæ
Help	Cancel OK	Help Cancel OK	Help Cancel OK

Gnumeric displays first the conditional characteristics of the dependent variable, then it provides the ANOVA table.

🐻 credit_ap	🤨 credit_approval.xlsx - Gnumeric													
<u>E</u> ile <u>E</u> dit	<u>Eile E</u> dit <u>V</u> iew Insert Format <u>T</u> ools <u>S</u> tatistics <u>D</u> ata <u>H</u> elp													
🗋 🖻 🖬 🚔 🖧 💥 🕼 🔯 🦘 👻 💉 🌚 🗵 fixi 🦕 🐝 🖍 100%														
Candara 10 🔺 🔺 📄 🗄 🗐 🎟 🔁 🔠 🐯 % · ‰ 🔐 📰 🖅 🖓 🗛 🖽 👻 🥠 👻 🗸														
F14	F14 🏟 💥 🖑 👻 = =devsq('ex.10 - anova'!\$A\$2:\$A\$15,'ex.10 - anova'!\$B\$2:\$B\$15,'ex.10 - anova'!\$C\$2:\$C\$15)													
1	Furniture	HiFi	HouseHold		Anova: Single Factor	r								*
2	25	26	35											
3	27	28	40		SUMMARY									
4] 30	30	65		Groups	Count	Sum	Average	Variance					
5	31	30			Furniture	14	550	39.2857	91.4505					
6	35	34			HiFi	13	505	38.8462	93.3077					
7	36	36			HouseHold	3	140	46.6667	258.3333					
8	36	37												Ξ
9	37	41												
10	44	43			ANOVA									
11	45	43			Source of Variation	55	df	MS	F	P-value	F critical			
12	47	46			Between Groups	156.9505	2	78.4753	0.749972	0.481966	3.3541			
13	50	55			Within Groups	2825.2161	27	104.6376						
14	53	56			Total	2982.1667	29							
15	54													-
10	4												 - F	
∢ ex.9	- indep nor	n paramet	tric ex.	10	- anova 🕨					Sum =	= 2982.16	67		

The function DEVSQ is essential in the calculations.

Results											
Attribute_Y	Attribute_X		Statistical test								
age	reason	Value	Examples	Average	Std-dev	Variance decomposition					
		Furniture	14	39.2857	9.5630	Source	Sum of square	d.f.			
		HiFi	13	38.8462	9.6596	BSS	156.9505	2			
		HouseHold	3	46.6667	16.0728	WSS	2825.2161	27			
		All	30	39.8333	10.1407	TSS	2982.1667	29			
							Significance level				
						Statistics	Value	Proba			
						Fisher's F	0.749972	0.481966			

The results are consistent with those of Tanagra.

3.11 Other statistical analysis

Gnumeric provides other statistical analysis. We have a description of the available approaches on the online manual (see "Statistical Analysis").

4 Conclusion

A spreadsheet is not specifically a statistical and data mining tool. But nonetheless, because of its skills and abilities, it is widely used in the context of statistical data processing. One usual solution is to use add-ins (for Excel, Libre and Open Office). They allow to overcome the poorness of its mathematical and statistical functions in this context. Some of them are free. "XNUMBERS" package for instance is highly accurate (De Levie, 2008).

In this tutorial, we describe the Gnumeric spreadsheet. It is a viable alternative to "Excel / LibreOffice / OpenOffice + Add-in" solution. It is a lightweight, multi-platform standalone tool that has all the necessary skills in handling and preparing data. It incorporates various statistical methods absent from the traditional spreadsheets. The Gnumeric's developers cooperates with those of R Software in order to improve the accuracy of the procedures (http://en.wikipedia.org/wiki/Gnumeric). We observe that the statistical functions are effective and provide valid results. Definitely, the computational library will improve positively over the years, Gnumeric is certainly a tool with potential for the future.

5 References

R. De Levie, « Advanced Excel for scientific data analysis », Oxford University Press, 2008.

K.B. Keeling, R. Pavur, « <u>Statistical Accuracy of Spreadsheet Software</u> », The Amercial Statistician, 65:4, 265-273, 2011. This paper is interesting because it proposes a particularly clear approach to evaluate statistical tools outputs, based on data and results provided by the NIST (<u>Statistical</u> Reference Datasets – National Institute of Standard and Technology).

Dana Lee Ling, « Introduction to Statistics Using LibreOffice.org Calc, Apache OpenOffice.org Calc and Gnumeric – Statistics using open source software », Edition 5.2, 2012; http://www.comfsm.fm/~dleeling/statistics/text5.html

B.D. McCullough, « Fixing Statistical Errors in Spreadsheet Software: The cases of Gnumeric and Excel », in Computational Statistics and Data Analysis Statistical Software Newsletter, 2004; http://www.csdassn.org/software_reports/gnumeric.pdf.

Gnumeric, « The Gnumeric manual, version 1.12 ».

Wikipedia, « Comparison of spreadsheet software ».