# 1 Objectif

#### Présentation du tableur GNUMERIC.

Le tableur est un outil privilégié des data scientist. C'est ce que nous révèle l'enquête annuelle du portail <u>KDNuggets</u>. Excel arrive régulièrement parmi les trois logiciels les plus utilisés ces cinq dernières années (enquêtes <u>2013</u>, <u>2012</u>, <u>2011</u>, <u>2010</u>, <u>2009</u>). En France, cette popularité est largement confirmée par sa présence quasi-systématique dans les offres d'emploi relatives au traitement exploratoire des données (statistique, data mining, data science, big data / data analytics, etc.) accessibles sur le site de l'<u>APEC</u> (Association Pour l'Emploi des Cadres). Excel est nommément cité, mais il faut surtout y voir une reconnaissance des aptitudes et capacités de l'outil tableur. D'autres suites bureautiques, dont certaines sont libres, proposent un module équivalent (ex. CALC de la suite LibreOffice).

Ce succès n'est pas très étonnant à vrai dire. Le tableur est très simple à utiliser, il possède des fonctionnalités multiples, y compris celle de savoir manipuler des tableaux de données pouvant aller jusqu'à 1.048.575 observations et 16.384 variables (ça nous laisse un peu de marge). Il est très largement répandu, tout le monde sait plus ou moins le manipuler. Pourtant, les informaticiens et statisticiens le considèrent parfois avec défiance. Certains ne sont pas tendres du tout (ex. « <u>The Risks of Using Spreadsheets for Statistical Analysis</u> », IBM SPSS Statistics ; un éditeur de logiciel statistique comme par hasard). C'est aller un peu vite en besogne je trouve. Il ne faut pas oublier qu'Excel n'a pas été spécifiquement conçu pour réaliser des calculs statistiques. Il n'est pas très équitable de le juger exclusivement à cette aune. Simplement, il nous importe de délimiter clairement ce qu'il est capable de faire.

Justement, Excel est largement utilisé certes, mais rarement en solo. Comme l'indiquent les enquêtes de KDNuggets, il est exploité en association avec des logiciels spécifiquement data mining qui présentent la précision voulue. On distingue parfaitement le partage des rôles dans cette optique : la préparation et le prétraitement des données sont réalisés sous tableur ; les traitements statistiques sont réalisés à l'aide des outils spécialisés. Ainsi a-t-on vu des éditeurs de logiciels proposer des extensions (add-ins, add-ons, packages, macros complémentaires) qui installent un menu additionnel et/ou des fonctions dédiés au traitement statistique et au data mining dans les tableurs. Même <u>SAS</u> s'y est mis, idem pour <u>Microsoft</u>. Il est indéniable d'ailleurs que l'utilisation des logiciels SIPINA et TANAGRA a été largement favorisée par les macros complémentaires facilitant l'échange des données avec les tableurs Excel et Libre / Open Office.

Ce tutoriel est consacré au tableur libre **Gnumeric** (<u>http://www.gnumeric.org/</u>). Il présente des caractéristiques intéressantes : le setup et l'installation sont de taille réduite parce qu'il ne fait pas partie d'une suite bureautique ; il est rapide et léger<sup>1</sup> ; il est dédié au calcul numérique et intègre de manière native un menu « Statistics » avec les procédures statistiques courantes (tests paramétriques, tests non paramétriques, régression, analyse en composantes principales, etc.) ; et, il semble plus précis que les tableurs de référence (McCullough, 2004 ; Keeling and Pavur, 2011). Ces deux derniers aspects ont attiré mon attention et m'ont convaincu de l'étudier plus en détail. Dans ce qui suit, nous effectuons un rapide tour d'horizon d'une sélection des procédures statistiques de Gnumeric. Pour certaines, nous comparons les résultats à ceux de **Tanagra 1.4.50**.

# 2 Données

Le fichier « credit\_approval.xlsx » décrit n = 30 individus demandeurs de crédit.

🐻 credit_ap	proval.xlsx - Gnu	umeric				_	-				2
File Edit V	View Insert	Format Too	ls Statistics	Data Help							
			Descri	ptive Statisti	cs	► z	<li>100 10</li>	0% 👻			
			Sampl	ing			IA   🛄 💴				
Arial 10	AA	A	Depen	dent Observ	ations		(	Až A <u>z</u>	<b>▼</b> 🗄	• A •	
			Two Sa	ample Tests							
C16	4	X 🖑	<ul> <li>Multipl</li> </ul>	e Sample Te	sts	۱.					
	A	В			E	F	6				
	reason	guarantee	Insurance	male.wage	temale.wage	Inc.nousenoid	Tamily.size	Inc.per.nead	age	acceptation	
		yes	yes	1230	1021	2259	2	2060	31	no	-
	HIFI	yes	yes	2398	1740	4138	2	2069	43	yes	-
4	Furniture	no	yes	1941	1220	3109	2	1004	54	yes	-
	Furniture	yes	yes	1740	1579	3319	4	030	30	yes	-
		yes	yes	1926	1420	335Z		1510	37	yes	-
		yes	yes	1370	1000	3031	2	1010	20	yes	-
		yes	yes	2230	1674	2001	2 E	706	30	yes	-111
10	Euroituro	yes	yes	2307	2154	3301		1099	41	yes	-
10	Euroituro	yes	yes	2230	2134	4330	- 4	1030	40	yes	-111
12	Eurniture	yes	yes	977	2000	2527	A	632	25	yes	-
12	Eurniture	yes	110	1666	1400	2027	4	740	20	110	-
14	Eurniture	yes	yes	1361	1400	2000	- 4	977	53	yes	-
15		yes	yes	1500	996	2332	5	479	46	yes	-
16	HiFi	yes	yes vec	2600	3107	5707		1/07	40	yes	-
17	HIEI	yes	<u>vec</u>	2000	2833	5/07	4	1358	30	yes	
18	HiFi	yes	yes	1799	1/96	3295		659	36	yes	-
19	HouseHold	yes	Vec	2540	1335	3875	4	939	40	yes	
20	Eurniture	yee	- 303 - 203	1909	1178	3087	3	1029	40		-111
21	Furniture	JUS VPS	Ves	2976	1753	4729	4	1182	36	ves	
22	HiFi	Ves	Ves	947	1226	2173	2	1086	56		
23	Furniture	ves	ves	1442	734	2176	3	725	27	ves	
24	HouseHold	ves	ves	834	1399	2233	4	558	35	ves	
25	Furniture	ves	ves	1063	1257	2320	2	1160	36	no	
26	HiFi	ves	ves	2266	1499	3765	2	1882	55	Ves	
27	HiFi	ves	ves	1127	1661	2788	4	697	37	no	
28	HiFi	ves	ves	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	ýes	γes	1229	2000	3229	4	807	43	yes	-
	<									+	
dataset								Sum	= 0		

Figure 1 - Fenêtre principale de Gnumeric avec le menu "Statistics"

<sup>&</sup>lt;sup>1</sup> Dixit la documentation. Il démarre rapidement, il n'y a aucun doute là-dessus. Le bilan est moins flatteur quand il s'agit de traiter les données. Un rapide test sur la durée d'importation et l'occupation mémoire d'un fichier texte avec séparateur tabulation de 500.000 observations et 22 variables a fourni les résultats suivants : Excel, 14 sec., 131 Mo ; Libre Office Calc, 180 sec., 256 Mo ; Gnumeric, 45 sec., 766 Mo. La question mérite d'être creusée en tous les cas.

Nous disposons de p = 9 variables (5 quantitatives, 4 qualitatives) : reason (motif de la demande), guarantee (existence d'une garantie), insurance (assurance), male.wage (salaire du demandeur), female.wage (salaire de sa conjointe), inc.household (revenus du ménage, formée par l'addition des deux salaires), family.size (nombre de personnes dans le ménage), inc.per.head (revenu par tête = revenu / nombre de personnes ; age (âge du demandeur de crédit), acceptation (décision de l'établissement prêteur).

Nous utilisons la version 1.12.12 pour Windows dans ce tutoriel, avec les menus en anglais. Une variante Linux est disponible. L'interface et le mode opératoire sont identiques.

Book1.g	numeric - (	Gnumer	ic						Fr	] t₊ ⊠	<b>■)) 1</b> 6	:10 🖞
			8	3 🖁			• 冷	-	$\sum$	f(x)	: 17	*
	Sans 10	a	a		-	<u>+a</u> + 	⊒ 88	Ģ	%•	و. و. <del>4</del>		*
	A1		•	8 🗸	• =							
		A )	В	С	D	E	F	G	H		J	K
	2											
	5											
	7											
	10											
	12 13											
A	15											
	17 18											
ġ	20											
	22											
	◀ Feuille	1 Feu	ille2 🕨						Som	ime = 0		

Figure 2 - Gnumeric sous Ubuntu

Dans les sections suivantes, nous décrivons plusieurs procédures statistiques de Gnumeric, avec toujours le même schéma : comment organiser les données pour pouvoir réaliser les traitements ; comment lancer et paramétrer l'outil ; comment lire les résultats.

# 3 Traitements statistiques sous Gnumeric

## 3.1 Statistiques descriptives

Nous désirons calculer les statistiques descriptives pour les variables quantitatives. Nous copions celles-ci dans une nouvelle feuille « ex.1 – desc. Stat », nous sélectionnons les données, puis nous actionnons le menu Statistics / Descriptive S

🐻 credit_app	oroval.xlsx -	Gnumeric		-	and states		
File Edit \	/iew Inse	rt Format	Tools St	tatistics	Data Help		
				Descrip Samplii	tive Statisti ng	cs	Correlation Covariance
Arial 10	AA	Α		Depend One Sa	lent Observ mple Tests	ations	Descriptive Statistics
F31		💫 🐰	4	Two Sa Multiple	mple Tests Sample Te	sts	Ranks And Percentiles
	Α	B	С	D	E	F	G H I J K L
	male.wage	emale.wag	c.househo	family.siz	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	
10	2236	2154	4390	4	1098	45	Grouped by: 💿 Columns 🚊 🛛 🥂
	3492	2088	5580	2	2790	44	
12	927	1600	2527	4	632	25	🗌 🔿 Rows
13	1566	1400	2966	4	742	35	
14	1361	1571	2932	3	977	53	O Areas
15	1500	896	2396	5	479	46	
<u> </u>	2600	3107	5707	4	1427	30	
17	2600	2833	5433	4	1358	30	
18	1799	1496	3295	5	659	36	
19	2540	1335	3875	4	969	40	
20	1909	1178	3087	3	1029	47	
	2976	1753	4729	4	1182	36	
	947	1226	2173	2	1086	56	
23	1442	734	2176	3	725	27	
24	834	1399	2233	4	558	35	
25	1063	1257	2320	2	1160	36	
26	2266	1499	3765	2	1882	55	
	1127	1661	2/88	4	697	37	
28	1425	1001	2426	3	809	26	Heip Cancel OK
29	//8	964	1742	2	871	65	
30	459	480	939	2	4/0	34	
31	1229	2000	3229	4	807	4j	
37	•	I	1	1	1	I	
dataset ex	(.1 - des	c. stat.					Sum = 229519.06666666667

Dans l'onglet INPUT, la plage de données doit être correctement sélectionnée ; elle est organisée en colonnes ; il est très important de cocher l'option « Labels » pour signifier au logiciel que la première ligne correspond aux noms des variables.

Nous ne modifions rien dans l'onglet STATISTICS, dans OUTPUT nous spécifions la localisation des sorties. Notons une option « Enter into cells : Formulae ». Il signifie que les résultats seront insérés sous forme de formules. De fait, si des valeurs de la plage de données sont modifiées, les résultats seront automatiquement mis à jour. Cette propriété est particulièrement intéressante. En l'état, elle ne permet cependant pas de s'adapter automatiquement à une modification des dimensions des données (nombre de lignes et de colonnes) mais, avec un peu de travail, ça doit pouvoir se faire.



Nous obtenons, entres autres, la moyenne, la médiane, l'écart-type, etc. (les résultats ont été formatés pour rendre la lecture plus facile).

🍓 *credit_ap	proval.xlsx -	Gnumeric	1 C				1000	S. 16.	100		Manda		- 0 X
File Edit \	√iew Insei	rt Format	Tools St	tatistics D	ata Help								
			( G 🕻	1 🤦	<b>-</b> 🤶	- 😜	$\sum f(x) = \frac{1}{2} \frac{z}{2}$	. 🗎 100	)% 🔻				
Candara	11					=   (SA) <	©∠ <b>. 49</b> .0 .00			□ <b>-</b> \$	- A -	-	
canadra	1						/U - 100 G.O			ш • •,			
13	<b>,</b>	s) 🐰	4 🗸	<b>=</b> =sq	rt(var('ex.:	1 - desc. st	at.'!\$A\$2:\$A\$31)/	count('e×.1 -	desc.stat.'!\$	A\$2:\$A\$31))			
	A	В	C	[ D ]	E	F	q н	<b>I</b>	J	K	L	[ M ]	N
1	male.wage	emale.wag	c.househo	family.size	nc.per.hea	age		male.wage	female.wage	inc.househola	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							
19	2540	1335	3875	4	969	40							
20	1909	1178	3087	3	1029	47							
21	2976	1753	4/29	4	1182	36							
22	947	724	2173	2	1086	55							
23	1442	1399	2170	3	7 20 558	- 27 - 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							
29	778	964	1742	2	871	65							
30	459	480	939	2	470	34							-
	∢												F.
dataset ex	x.1 - des	c. stat.									Sum =	130.9361	
			_						-				

Attardons-nous un instant sur l'écart-type de la moyenne de « X : male.wage ». En cellule I3, nous distinguons la formule  $s_{\bar{x}} = \sqrt{\frac{s_x^2}{n}} = \sqrt{\frac{514328.1885}{30}} = 130.9361$ . La variance estimée  $s_x^2$  de X est en cellule I7.

Par comparaison, nous obtenons les résultats suivants pour « male.wage » sous Tanagra. Les résultats concordent en tous points.

TANAGRA 1.4.50 - [More Univariate	cont stat 1]						x
<u>File D</u> iagram Component <u>W</u> ir	ndow <u>H</u> elp					- 6	5 ×
🗅 📽 🔚 🛛 👪							
Default title			Res	sults			*
📃 🔠 Dataset (credit_approval_data	aset.xls)	Attribute		Sta	ats		
E Define status 1				Stati	stics		
More Univariate cont	stat 1		Average		1741.1333		
			Median		1653.0000		
			Std dev. [Coef of var	riation]	717.16	68 [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 * 2266.0	0 [1037.00]	
		-	Skewness (std-dev)			27 (0.4269)	
			Kurtosis (std-dev)		-0.23	80 (0.8327)	-
		•					
		Componen	ts				
Data visualization	Statistics	Nonparar	netric statistics	Insta	nce selection		
Feature construction Feat	ure selection	Re	gression	Fac	torial analysis		
PLS C	Clustering	Spy	learning	Met	a-spv learning		
Spv learning assessment	Scoring	As	sociation				
ANOVA Randomized Blocks	Box's M Test		🛱 Fisher's test		<b>E</b> (	Group explo	orati
🛱 Bartlett's test 🖉	Brown - Forsy	the's test	🚺 Group chara	cterizat	ion 🛃	Hotelling's T	2
· · · · · · · · · · · · · · · · · · ·							•

## 3.2 Matrice des corrélations

Nous reprenons les mêmes données pour calculer la matrice des corrélations. Nous les dupliquons dans une nouvelle feuille « **ex.2 – corr matrix** ». Après avoir sélectionné la plage de valeurs, nous actionnons le menu **Statistics / Descriptive Statistics / Correlation**. De nouveau, nous vérifions la plage d'entrée, indiquons que la première ligne correspond à des noms de variables, indiquons les coordonnées de la plage de sortie.

Correlation		Correlation
Input Lt		Input Output
Input range: Grouped by:	<ul> <li>'ex.2 - corr matrix'!\$A\$1:\$F\$31</li> <li>Columns</li> <li>Rows</li> </ul>	Output Placement  New sheet  New workbook  Output range: lex 2 - corr matrix!!!!
☑ Labels	⊘ Areas	Output Formatting     Autofit columns     Output range     Retain output range formatting     Retain output range comments     Enter into cells:     Formulæ
Help	Cancel OK	<u>H</u> elp <u>Cancel OK</u>

Les corrélations sont obtenues à partir de la formule CORREL.

👸 credit_app	proval.xlsx - Gn	umeric								x	
<u>E</u> ile <u>E</u> dit <u>)</u>	<u>√</u> iew <u>I</u> nsert	F <u>o</u> rmat <u>T</u> oo	ols <u>S</u> tatistic	s <u>D</u> ata <u>H</u> el	р						
		l 🔏 🖣	ē 🖻 🚶	5 🔻 🦿	▼ 😜	$\sum f(x) =$	z Na	100%	•		
Candara	Candara 10 🔺 🔺 📄 📰 💭 🌐 🖽 🖽 💭 % · ‰ 🔐 🛛 📰 🕅 🗛 📔 👻										
J4	4 🌍 💥 🛫 👻 🚽 a correl('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			
8				•							
9											
14											
14											
	4									•	
	x.2 - corr m	atrix					Sum = 0.	8570			

## 3.3 Analyse en composantes principales

Nous créons une troisième feuille « **ex.3 – PCA** » et copions les variables quantitatives. Nous sélectionnons les données puis actionnons le menu **Statistics / Dependent Observations / Principal Components Analysis**. Voici les paramètres associés :

Principal Components Analysis	Principal Components Analysis
Input Lt	Input Output
Input range: Vex.3 - PCA'!\$A\$1:\$F\$31	Output Placement
Grouped by:	<ul> <li>New sneet</li> <li>New workbook</li> </ul>
© <u>R</u> ows © <u>A</u> reas	◎ Output range: 'ex.3 - PCA'!H1
☑ <u>L</u> abels	Output Formatting  Autofit columns
	🗹 Clear output range
	Retain output range formatting
	Retain output range comments
	Enter into cells: Formulæ
<u>H</u> elp <u>C</u> ancel <u>O</u> K	Help Cancel OK

Nous obtenons :

🤠 credit_ap	proval.xlsx - Gr	numeric	-	-	_			x			
File Edit	View Insert	Format To	ols Statisti	cs Data He	elp						
			G D	🦕 🚽 🤞		$f(\mathbf{x})$	A Z	•			
		<b>4</b> 8		2 · v	·   🥣		Sala a sa				
Candara	10 🔺	AA		<b>VP D</b>	88 🖗	% · 4	00.00	•			
114	4	8 4	• =	{=eigen(I10	/(110-1)*13:1	48)}(6,7)[0]	[0]				
1	Principal Com	ponents Anal	vsis					-			
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	]	ξι	ξ2	<b>ξ</b>	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.househol	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%				
	•							•			
◀ ex.	3 <b>-</b> PCA				Sum = 204	9615.200		//			

Nous distinguons successivement :

- La matrice des covariances, Gnumeric effectue une ACP non normée ;
- Le nombre d'observations par variable, les moyennes et les variances estimées ;
- Les valeurs propres par axe factoriel ;
- Les vecteurs propres ;
- Les corrélations des variables avec les axes ;
- La proportion d'inertie restituée par les axes.

Il n'y pas d'options pour réaliser une ACP normée. Une solution simple consisterait à remplacer manuellement les valeurs des covariances par les corrélations (substituer CORREL à COVAR dans la feuille de calcul). Les autres formules sont à conserver tels quels. Les résultats sont automatiquement mis à jour. Je trouve cette potentialité assez enthousiasmante.

En rapprochant les sorties avec celles de Tanagra, j'ai constaté une différence au niveau des valeurs propres (ci-dessous le tableau des valeurs propres de Tanagra) :

Axic		Difference	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	-	-	-

Nous trouvons l'explication de cet écart dans la formule utilisée par Gnumeric (la cellule incriminée et la formule sont indiquées par des flèches dans la copie d'écran ci-dessus). Gnumeric affiche  $\frac{n}{n-1} \times \lambda_1 = \frac{30}{29} \times 1981294.694 = 2049615.2$  où n = 30 est le nombre d'observations,  $\lambda_1$  est la première valeur propre de la matrice de covariance, affichée par Tanagra. Les vecteurs propres sont pondérés de la même manière. La correction est mineure pour nous si nous souhaitons retrouver des résultats conformes aux ouvrages de référence en français. Les corrélations des variables avec les axes factoriels, essentielles pour l'interprétation, ne sont pas affectées par cette pondération.

## 3.4 Régression linéaire

Mettons que nous souhaitons expliquer la taille des familles à partir du revenu des ménages et de l'âge des demandeurs de crédit (ok, l'exemple est un peu loufoque, mais notre objectif est de décrire le fonctionnement de Gnumeric, pas de faire un cours d'économétrie). Nous copions dans une nouvelle feuille « **ex.4 – Linear reg** », en les mettant dans l'ordre, les

variables « inc.household », « age » et « family.size ». Nous actionnons le menu Statistics / Dependent Observations / Regression. Nous spécifions les paramètres suivants.

Regression X	Regression	Regression
Input Output	Input Options ut	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	,	<ul> <li>New workbook</li> </ul>
	Calculate residuals	Output range: 'ex.4 - Linear reg'!E1
X variables: 'ex.4 - Linear reg'!\$A\$1:\$B\$31		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
l		
Help Cancel OK	Help Cancel OK	Help Cancel OK

Y est la variable dépendante (familiy.size), X représente le bloc des variables indépedantes (inc.household et age). Gnumeric s'appuie essentiellement sur la fonction LINEST. Il réorganise les résultats pour une présentation plus conforme aux sorties des logiciels de statistique. Il pioche les différentes valeurs dans un tableau interne avec la fonction INDEX.

诸 *credit	_approval.xlsx - Gr	numeric	-	_		_	-		_				×
File Edit	: View Insert	Format To	ols Statistics	Data Help									
				_ >			Z Aa	10.00/					
		Q   🚜	2 🖬   🤈	• 🦿	▼   🥹 .	∑f(x) <u>v</u> iz	: 🖬   🛄	100%	•				
Candar	a 10 🔺 📝	AA			E 🐶 🕺	• 👫	.00	*= A	Az	₩ ▼	oh <b>▼</b> ,	A 🗸	
E17	~	· ·		in day (line at	(Inc. 4. Line	an na alltacha	utotal lave (	1 1		.+D+01 TD		1 2))(1 1)	10101
F1/	<b>V</b>		• = {=	index(linest	('ex.4 - Line	ar reg 1\$C\$2	::\$C\$31,'eX.4	+ - Linear	reg!!\$A\$2	:\$8\$31,18	UE, IRUE),	1,3)}(1,1	[0][0]
	E	F	G	н	1		K	L	M	N	0	P	
1	SUMMARY OU	TPUT	Response Varia	family.size			,						
2													
3	Regression Sta	atistics											
4	Multiple R	0.397913											
5	R^2	0.158335											
6	Standard Erro	1.011171											
7	Adjusted R^2	0.095990											=
8	Observations	30											
9													
10	ANOVA												
11		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										
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	inchousehold	0.000211	0.000166	1.264897	0.216717	-0.000131	0.000552						
- 19	age	-0.033300	0.018542	-1.795865	0.083716	-0.071345	0.004746						<u> </u>
20	4												
<b>∢</b> ∈	x.3 - PCA	ex.4 -	Linear reg							Sum = 3.8	342254		

Nous observons ainsi successivement les résultats généraux (R2, etc.), le tableau d'analyse de variance (avec le F de Fisher pour le test de significativité globale), le tableau des coefficients (avec les tests de significativité et les intervalles de confiance).

Nous retrouvons les mêmes résultats avec Tanagra, l'organisation est identique.

TANAGRA 1.4.50 - [Multiple linear regression 1]							×
💇 <u>F</u> ile <u>D</u> iagram Component <u>W</u> indow <u>H</u> elp						- 5	×
		_					
Default title	Report	(X'X)^(-1) ma	ıtrix				
□-□ Dataset (credit_approval_dataset.xls) □ 🚼 Define status 1	Global re	sults					*
More Univariate cont stat 1	Endogenous att	ribute		family.size			
Principal Component Analysis 1	Examples			30			
🖶 🚰 Define status 2	R <sup>2</sup>			0.158335			
🔣 Multiple linear regression 1	Adjusted-R <sup>2</sup>			0.095990			
1	Sigma error			1.011171			
	F-Test (2,27)		2.539	6 (0.097584)			=
	Analysis o	of varianc	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					
	Attrib	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	_
	•					•	
<u> </u>	Co	mponents					
Data visualization Statistics	Nonparam	etric statistics	Instance	e selection	Feature o	onstruction	
Feature selection Regression	Factor	ial analysis	F	PLS	Clus	tering	
Spv learning Meta-spv learning	Spv learni	ng assessment	Sc	oring	Asso	ciation	
🛃 Backward Elimination Reg 🖄 DfBetas	🛃 Forv	vard Entry Reg	ression 🛛 🖾 Nu	SVR	Ŷ⊴ty R	egression Asse	ssm
🚓 C-RT Regression tree 🛛 🔀 Espilon SVR	🛃 Mul	tiple linear reg	ression 🛛 🖄 Ou	tlier Detection	ه <mark>ُ،</mark> R	egression tree	
·							•
							đ

## 3.5 Comparaison d'une moyenne à un standard

Nous souhaitons savoir si l'homme et la femme d'un même ménage ont des salaires comparables. Pour ce faire, nous copions les deux colonnes dans la feuille « **ex.5 – One sample t-test** ». Nous créons nouvelle variable DIF composée de l'écart (male.wage - female.wage). Sous l'hypothèse nulle, les salaires sont identiques, cette différence devrait être égale à 0 en moyenne. Nous effectuons donc un test de comparaison à un standard.

Après avoir sélectionné la colonne DIF, nous actionnons le menu **Statistics / One Sample Tests / Claims About a Mean**. Nous paramétrons la procédure comme suit.

### Tanagra

### Ricco Rakotomalala

🐻 Claims About a Mean	Claims About a Mean	Claims About a Mean
Input Output Input ange: One sample t-test'!\$C\$1:\$C\$31 Grouped by:  Columns Rows Areas Labels	Input Test, put Predicted Mean: 0 Alpha: 0.05	Input Test       Output Placement         Output Placement       New sheet         New workbook       Output range: One sample t-test'IEI         Output Formatting       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

Au risque 5%, nous rejetons l'hypothèse de la nullité (Predicted Mean = 0 dans l'onglet TEST) de la moyenne de DIF<sup>2</sup>. Nous observons dans la cellule F6 la formule ( $\mu_0$  = 0 dans notre exemple) :

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

La p-value est obtenue à l'aide de la fonction TDIST.

🍓 *credit_ap	🔁 *credit_approval.xlsx - Gnumeric							
File Edit View Insert Format Tools Statistics Data Help								
🗋 🖻 🖥 🚔 🗸 🖌 🕼 🖬 🇯 🔸 🔹 🔶 🔹							•	
Candara 12 🔺 🔺 🔺 🖹 🗐 🗐 🖽 🖽 🔷 🌮 🐯 % · 比 🚅 📔 🖛							•	
F6	4	💥 🖑	• =	{=	(F3-F4)/sqrt(F5/F2)}(1,1)	[0][0]		
	A	В	С	D	)[E	[ <b>F</b> ]		
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	1 4500	4.400	100				F	
ex.5 - One	e sample t-	test			Sum = 2.13	06	1	

<sup>2</sup> Une autre piste consisterait à calculer le ratio entre les salaires et de comparer la moyenne à 1. La conclusion est la même, mais les valeurs de la statistique de test et de la p-value sont différentes [ex.5 (bis) – One sample t-test].

## 3.6 Comparaisons de moyennes – Echantillons appariés

Une autre manière de répondre à la question d'égalité des salaires intra-ménage consiste à réaliser un test de comparaison de moyennes pour échantillons appariés<sup>3</sup>. Nous copions les deux colonnes de salaires dans une nouvelle feuille « **ex.6 – Paired t-test** ». Nous actionnons le menu **Statistics / Two Sample Tests / Claims About Two Means / Paired Samples**.



Nous devons explicitement sélectionner les variables dans l'onglet INPUT, sans oublier de spécifier que la première ligne correspond au nom de variables (Labels). Les deux colonnes doivent obligatoirement la même longueur, sinon l'appariement n'a pas sens.

<sup>&</sup>lt;sup>3</sup> R. Rakotomalala, « Comparaison de populations - Tests paramétriques », version 1.2, Juin 2013 ; <u>http://eric.univ-</u>lyon2.fr/~ricco/cours/Comp\_Pop\_Tests\_Parametriques.pdf

🤠 credit_app	proval.xlsx - G	inumeric	-		1 . mar		x		
File Edit 🕚	File Edit View Insert Format Tools Statistics Data Help								
	🗋 📄 🔚 📇 🕺 😳 🔯 👆 👻 🦿 🕐 👻 🕼 🎼 🗤 🔽 🔽								
Candara	Candara   12 🖪 🖪 🖪 🗄 🖽 📼 🎷 🐯 % • 號 🔐   🐺 🖅 🔻								
E10	Ę	> 🗶 🦿		= =(E7-E6)/(E8/(E9+1))^	0.5				
	A	В	С	D	[ <b>E</b> ]	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	=		
5	1740	1579		Pearson Correlation	0.5837				
6	1926	1426		Hypothesized Mean Difference	0				
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		-		
	•						•		
▲ ex	x.6 - Paire	d t-test			Sum = 2.130	6	//		

Par un procédé différent, nous obtenons strictement (les valeurs et distributions des statistiques de test sont identiques) le même résultat que précédemment (section 3.5). Les hommes et les femmes ont des niveaux de salaire différents à l'intérieur des ménages.

## 3.7 Test non-paramétrique – Echantillons appariés

Nous pouvons répondre à la même question, comparer les salaires des hommes et des femmes à l'intérieur des ménages, en utilisant le test des rangs signés de Wilcoxon. La statistique de test est basée sur le rang des écarts, et non sur leur amplitude<sup>4</sup>. L'intérêt de cette procédure est que l'on s'affranchit de l'hypothèse de normalité des distributions.

Nous copions les colonnes des salaires dans la feuille « ex.7 – Paired Wilcoxon ». Nous actionnons le menu Statistics / Two Sample Tests / Claims About Two Medians / Wilcoxon Signed Rak Test.

<sup>&</sup>lt;sup>4</sup> R. Rakotomalala, « Comparaisons de populations – Tests non paramétriques », Août 2008 ; <u>http://eric.univ-</u>lyon2.fr/~ricco/cours/Comp\_Pop\_Tests\_Nonparametriques.pdf (pages 130 et suivantes).

### Tanagra

### Ricco Rakotomalala

Claims About Two Medians (Paired Samples)	Claims About Two Medians (Paired Samples)	Claims About Two Medians (Paired Samples)
Input	Input Test	Input Test Output
Variable 1 range: 'ex.7 - Paired Wilcoxon'!\$A\$1:\$A\$31	Sign Test  Vilcoxon Signed Rank Test Hypothesized difference of medians: 0	Output Placement New sheet New workbook
✓ Labels	Alpha: 0.05 🗕 🖶	Output range: 'ex.7 - Paired Wilcoxon'!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	Help Cancel OK

### Nous obtenons :

🥫 credit_app	proval.xlsx - G	numeric						- O <mark>X</mark>
File Edit View Insert Format Tools Statistics Data Help								
🗋 🖻 🖬   📇 🤮   🚜 🗊 🔯   🦻 💌 🥐 👻   🥹 🕱 f 🖉 🦕 🐝   🖍 100% 🔽								
Candara 12 🗛 🗛 📄 🗐 🗐 🖽 🖽 🖽 😽 🐺 😽 + 🏶 🔐 🛛 📰 🗮 🗮 🕶 🕬							• •	
E10	4	) 🐰 🖑	▼ = if(E5<12,#N/A,norn	ndist(E8+0.5,E5*(8	5+1)/4,sqrt(E	5*(E5+1)/	4*(2*E5+1	.)/6),TRUE))
	A	В	C D	) E	F	G	H	I [
1	male.wage	female.wage	Wilcoxon Signed Rank Test	male.wage	female.wage			-
2	1238	1021	Median	1653	1461			
3	2398	1740	Observed Median Difference	191.5				=
4	1941	1228	Predicted Median 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				
10	2236	2154	P(T≤t) one-tailed	0.035147				
11	3492	2088	P(T≤t) two-tailed	0.070294				
12	927	1600						
10	<	4.400					1 1	•
	Paired t-tes	t ex.7 - P	aired Wilcoxon 🕨		Sur	n = 0.035	147	

Les effectifs étant suffisamment élevés (n  $\ge$  12), Gnumeric fournit la p-value basée sur la loi normale en calculant à la volée la valeur Z. Pour un test bilatéral, nous avons p-value = 0.070294. L'écart de salaire est moins évident semble-t-il avec ce test.

Par rapport aux sorties de Tanagra, nous observons Z = 1.820298 (affiché explicitement cette fois-ci) avec une p-value =0.068714.

Attribute_Y		Attrib	oute_X	Statistical test		
male.	male.wage female.wage Me		Measure	Value		
Avg	1741.133333	Avg	Avg 1508.966667		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	

Les p-value divergent. Il faut savoir pourquoi. J'ai un peu creusé la question, la différence se joue au niveau de la correction de continuité. En calculant Z'

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

Avec la fonction de répartition de la loi normale centrée et réduite  $\Phi(.)$ , nous avons :

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

Exactement la valeur produite par Gnumeric.

#### 3.8 Test paramétrique – Echantillons indépendants

Notre propos est de comparer les revenus par tête des ménages selon l'acceptation de la demande de crédit. Il s'agit d'un test pour échantillons indépendants puisque la variable « acceptation » induit une partition non recouvrante des individus.

Ce test nécessite une mise en forme particulière des données dans Gnumeric dans la feuille « **ex.8 – indep parametric** ». Plutôt que la présentation initiale individus x variables, nous devons créer des colonnes de valeurs « inc.per.head » pour chaque modalité de « acceptation » (yes, no). Ces 2 colonnes n'ont pas forcément la même longueur puisque les effectifs conditionnels peuvent être différents.

🐻 credit_ap	📴 credit_approval.xlsx - Gnumeric							
File Edit	File Edit View Insert Format Tools Statistics Data Help							
lnc.p	er.he	ad for		- <u>-</u>	• 🕐	• 🛛 🕹	$\sum f(x)$	•
Acce	ptatic	n = n	0	Inc.	per.h	lead f	or	•
A12		n 🔊 🖓	T	Acc	ontat	tion -	MOG	
		В		ACC	epta	1011 –	yes	
		Vec					<u></u>	
2	1 470	479						
3	632	558						
4	697	659						
5	809	725						
6	871	742						
7	1029	796						
8	1086	807						
9	1130	830						E
10	) 1160	969						
[ 11	]	977						
12		1098						
13	ļ	1117						
14	Į	1182						
15	Į	1358						
16	1	1427						
1/	J	1516						
18	ļ	1584						
19	J	1773						
20	1	1002						
22	1	2009						+
23	1	2750						
23	•		· ·	III				•
indep par	ametric				Sum :	= 0		

Voici le paramétrage de la procédure **Statistics / Two Sample Tests / Claims about two means / Unpaired Samples, Unequal Variances**. Nous faisons l'hypothèse que les variances conditionnelles sont différentes.

Claims About Two Means	Cla	aims About Two Means		
Input Input Input	Input	t Populations		
Variable 1 range: 0.8 - indep parametric!!\$A	\$1:\$A\$10	riables are:	Paired	Onpaired
Variable 2 range: 0.8 - indep parametric!!\$B	\$1:\$B\$22	pulation variances are:	🔘 Known	Onknown
✓ Labels	Po	pulation variances are:	🔘 Equal	Onequal
Help Cancel		Help	Cancel	ок
Claims About Two Means		ims About Two Means		
Claims About Two Means		iims About Two Means	<b>þ</b>	
Claims About Two Means Input Populations Test Hypothesized mean difference: 0		ims About Two Means t Populations Test Output tput Placement	5	
Claims About Two Means Input Populations Test Ut		ims About Two Means t <u>Populations</u> <u>Test</u> Output t <b>put Placement</b> ) New <u>s</u> heet	5	
Claims About Two Means Input Populations Test Hypothesized mean difference: Alpha: 0.05		ims About Two Means t <u>Populations Test</u> Output tput Placement New <u>s</u> heet New <u>w</u> orkbook	5	
Claims About Two Means Input Populations Test Hypothesized mean difference: Alpha: 0.05		ims About Two Means t <u>Populations</u> <u>T</u> est <u>O</u> utput <b>tput Placement</b> New <u>s</u> heet New <u>w</u> orkbook Output <u>r</u> ange: 'ex.8 - indep	o parametric	"ID1
Claims About Two Means Input Populations Test ut Hypothesized mean difference: Alpha: 0.05		ims About Two Means t <u>Populations</u> <u>Test</u> <u>O</u> utput tput Placement New <u>s</u> heet New <u>w</u> orkbook Output <u>r</u> ange: <sup>1</sup> ex.8 - indep tput Formatting Autofit columns	o parametric	'!D1
Claims About Two Means		ims About Two Means t <u>Populations</u> <u>Test</u> <u>O</u> utput tput Placement New <u>s</u> heet New <u>workbook</u> Output <u>r</u> ange: <u>'ex.8 - indep</u> tput Formatting A <u>u</u> tofit columns <u>C</u> C <u>l</u> ear output range	o parametric	"ID1
Claims About Two Means		ims About Two Means t <u>Populations Test</u> Output tput Placement New <u>s</u> heet New <u>w</u> orkbook Output <u>r</u> ange: 'ex.8 - indep tput Formatting Autofit columns C Clear output range Retain output range formatt	p parametric	"D1
Claims About Two Means		ims About Two Means t <u>Populations</u> Test Qutput tput Placement New <u>s</u> heet New <u>workbook</u> Output <u>r</u> ange: <u>'ex.8 - indep tput Formatting Autofit columns C Clear output range Retain output range formatt Retain output range comme</u>	o parametric ing nts	"D1
Claims About Two Means		ims About Two Means t <u>Populations Test</u> Output tput Placement New <u>s</u> heet New <u>workbook</u> Output <u>r</u> ange: 'ex.8 - indep tput Formatting Autofit columns C Clear output range Retain output range formatt Retain output range comme nter into cells: Formulæ	o parametric ing nts	"ID1 R
Claims About Two Means Input Populations Test ut Hypothesized mean difference: 0 Alpha: 0.05 Help		ims About Two Means  t <u>Populations</u> <u>Test</u> <u>O</u> utput  tput Placement  New <u>sheet</u> New <u>workbook</u> Output <u>range</u> : <u>'ex.8 - indep  tput Formatting</u> Autofit columns Clear output range Retain output range formatt Retain output range comme nter into cells: Formulæ	o parametric ing nts	

La procédure s'appuie sur le test de Welch (<u>http://en.wikipedia.org/wiki/Welch's\_t\_test</u>). La statistique est relativement facile à calculer, elle suit une loi de Student sous l'hypothèse nulle (les moyennes sont identiques). Le véritable enjeu est dans le calcul des degrés de liberté. Gnumeric propose les résultats suivants, le degré de liberté est fractionnaire (df. = 27.99). Nous avons une p-value de 0.035396 pour un test bilatéral.

🐻 *credit_ap	🥐 *credit_approval.xlsx - Gnumeric							
File Edit \	File Edit View Insert Format Tools Statistics Data Help							
	🕒 🖻 🖥 📇 🗸   🔏 🗊 😰   🥠 🏹 🔻   🌚 🗵 f(x) 🦕 💑   🏦 100% 🔽							
Candara 12 🔺 🔺 📋 🗐 🗐 📰 🔛 🎦 🕄 🦃 % · ‰ 🔐   📻 😁   👻								
E7  (E3/E4+F3/F4)^2/((E3/E4)^2/(E4-1)+(F3/F4)^2/(F4-1))								
	A	В		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	0			
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						
<u> </u>		1358						
<u> </u>		1427						
17		1516					_	
18		1584					_	
19		1773			I			
20		1882						
21		2069						
		2790					-	
L	·						•	
	- indep	paramet	ric 🕩		Sum = 27.99		//	

Comparés à ceux de Tanagra, les résultats concordent excepté la p-value.

Attribute_Y	Attribute_X	Description					Statistical test
	Volue Examples Assess		Examples			-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		

Cette différence s'explique par la gestion des degrés de liberté fractionnaires : Tanagra utilise sur l'entier le plus proche (df = 28 dans notre cas) ; la fonction TDIST de Gnumeric semble s'appuyer sur une interpolation linéaire<sup>5</sup> (df = entre 27 et 28). L'écart est très minime quoiqu'il en soit. Il faut savoir l'expliquer simplement.

# 3.9 Test non paramétrique – Echantillons indépendants

Nous utilisons le test de Wilcoxon-Mann-Whitney pour le versant non paramétrique. Les données (feuille « **ex. 9 indep non parametric** ») doivent être organisés comme

<sup>&</sup>lt;sup>5</sup> http://fr.wikipedia.org/wiki/Interpolation\_linéaire

précédemment (section 3.8). Nous actionnons le menu Statistics / Two Sample Tests / Claims About Two Medians / Wilcoxon-Mann-Whitney test. Voici le paramétrage associé.

🕞 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 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

La statistique Z pour l'approximation normale n'est pas explicitement affichée, mais elle est utilisée pour le calcul de la p-value avec NORMDIST.

🗃 *credit_approval.xlsx - Gnumeric											
File Edit View Insert Format Tools Statistics Data Help											
🗋 🖻 🖥 🛔 🍇 🖡 🖻 😰 🥠 🕶 🦿 🚱 🗵 f(x) 🧞 💑 🛍 🛛 💌											
Candara 10 🔺 🔺 📄 🗐 🗐 🎟 🕶 🖼 🐯 % • ‰ 🔐 📻 💌 🗸											
E9 💫 💥 🛫 = =2*normdist(E8,E4*F4/2,sqrt(E4*F4*(E4+F4+1)/12),TRUE)											
	A	<u>В</u>	С	) D	<b>E</b>	F	G	н			j
1	no	yes		Wilcoxon-M	ann-Whitn	ey Test					
2	470	479			no	yes	Total				
3	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						_
10	1160	969									_
		977									_
12		1098									-
13		1117									_
14		1182									_
15		1358									_
16		1427									
		1516									
18		1584									
		1773									
20		1882									
21		2069									
		2790									-
	•	· ·									•
ex.9 - in	ex.9 - indep non parametric Sum = 0.16749										

	Value	Examples	Average	Rank sum	Rank mean	Mann-Whitney U	64
	no	9	875.8796	109	12.1111	E(U)	94.5
inc.per.headacceptation	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 propose exactement le même résultat, mais via une présentation différente.

# 3.10 Analyse de variance (ANOVA)

Nous souhaitons maintenant comparer l'âge des personnes selon le motif (reason) de la demande de crédit. Nous créons la feuille « **ex.10 – anova** ». Nous créons autant de listes de valeurs (âge) qu'il y a de modalités de « reason » (furniture, hifi et household). Les effectifs conditionnels ne sont pas forcément identiques.

Voici l'organisation des données dans la nouvelle feuille de calcul.

🐻 credit_app	proval.xlsx - G	numeric		-		• ×					
File Edit View Insert Format Tools Statistics Data Help											
				9.	• 🥐						
Value	s of age	accord	ding the I	evelso	of reas	son 🖵					
A1 Euroiture											
	A	B	C	D	<u> </u>	F					
	Furniture	HIFI	HouseHold								
	25	26 20	35			_					
	2/	20	40								
<u> </u>	30	30	60								
6	35	34									
7	36	36				_					
8	36	37									
9	37	41									
10	44	43									
11	45	43									
12	47	46									
13	50	55									
14	53	56									
<u>15</u>	54										
<u> </u>											
	•		111			•					
va			Sum :	= 0							

Nous actionnons le menu **Statistics / Multiple Sample Tests / ANOVA / One Factor** pour lancer la procédure. Dans la boîte de paramétrage, nous sélectionnons le bloc de données, qu'importe si certaines cellules sont vides.

### Tanagra

#### Ricco Rakotomalala

ANOVA - Single Factor	ANOVA - Single Factor	ANOVA - Single Factor
Input ons Output	Input Options but	Input Options Output
Input range:  'ex.10 - anova'!\$A\$1:\$C\$15	Alpha: 0.05 💳 🛟	Output Placement
Crouned by Columns		New sheet
Grouped by: Columns		New workbook
© Rows		Output range: 'ex.10 - anova'!E1
O Areas		Output Formatting
🖉 Labels		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 fournit les caractéristiques conditionnelles et le tableau d'analyse de variance.

🧓 credit_approval.xlsx - Gnumeric										8			
<u>Eile E</u> dit <u>V</u> iew Insert Format Tools <u>S</u> tatistics <u>D</u> ata <u>H</u> elp													
🗋 🔚 🔚 📇 🐘 🗊 😰 🖕 👻 🦿 😥 🌠 f(x) 🏪 💑 🚺 100% 🔽													
Candara 10 🗛 🗛 📄 🗐 📰 🔁 🔡 🦃 % · 🏶 🔐 📮 🖻 👫 🗛 🖿 💌 🔦 💌 👻													
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 Facto	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	M5	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												-
	4											1	
<b>∢</b> ex.9	Image: Construction of the second seco												

La fonction DEVSQ correspond à la somme des carrés des écarts. Elle joue un rôle important dans cette analyse. Les valeurs fournies sont raccords avec celles de Tanagra.

Results											
Attribute_X		Statistical test									
reason	Value	Examples	Average	Std-dev	Va	riance decompositio	on				
	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				
	Attribute_X	Attribute_X  Attribute_X  Value  Furniture HiFi HouseHold All	Attribute_X           Value         Examples           Furniture         14           HiFi         13           HouseHold         30	Results         Attribute_X       Description         Value       Examples       Average         Furniture       114       39.2857         HiFi       113       38.8462         HouseHold       30       39.8333         Reason       All       30	Results           Attribute_X         Description           Image: Attribute_X         Examples         Average         Std-dev           Image: Attribute_X         Image: Average         Std-dev         9.5630           Image: Attribute_X         Image: Average         Image: Average         9.5630           Image: Average         Image: Average         Image: Average	Attribute_X         Value         Examples         Average         Std-dev         Value           Image: Furniture         Image: Furniture<	Results         Attribute_X       Description       Statistical test         Attribute_X       Examples       Average       Std-dev       Statistical test         Image: Altribute_X       Furniture       Average       Std-dev       Source       Sum of square         Image: Altribute_X       Image: Altribute_X       Image: Altribute_X       Image: Altribute_X       Std-dev       Source       Sum of square         Image: Altribute_X       Image: Altri				

## 3.11 Autres techniques statistiques

Gnumeric propose d'autres techniques statistiques. Pour une description exhaustive, je conseille la lecture du manuel en ligne (chapitre « Analyse Statistique »).

# 4 Conclusion

Un tableur n'est pas un logiciel de statistique et de data mining en tant que tel. Je pense que tout le monde est d'accord avec cette idée. Il n'en reste pas moins que, de par ses qualités, il est très largement pratiqué dans les entreprises, y compris dans le cadre du traitement de données. Rien que pour cette raison, en tant que formateurs préparant les étudiants au monde professionnel, nous ne pouvons certainement pas passer à côté.

Pour palier la pauvreté des bibliothèques de fonctions mathématiques et statistiques des tableurs usuels (Excel, Calc de LibreOffice et OpenOffice), et disposer d'une précision de calcul satisfaisante, les add-ins constituent une réponse intéressante. Nous pouvons les télécharger et installer aisément. Nombre d'entre eux sont libres (<u>http://www.statsci.org/excel.html</u>). Certaines, comme la librairie « matrix.xla », sont très performantes et passent haut la main les tests de précision (<u>http://digilander.libero.it/foxes/</u>). Elles ouvrent la porte au calcul scientifique viable sous Excel (De Levie, 2008).

Dans ce tutoriel, nous avons décrit le logiciel Gnumeric. Il constitue une alternative au tandem « Excel / LibreOffice / OpenOffice + add-in ». C'est un outil standalone léger et multiplateforme qui dispose de toutes les aptitudes nécessaires en matière de manipulation et de préparation des données. Il intègre nativement plusieurs méthodes statistiques absentes des tableurs traditionnels. Les développeurs de Gnumeric coopèrent avec l'équipe de R afin d'améliorer la précision des procédures (<u>http://en.wikipedia.org/wiki/Gnumeric</u>). Nous constatons qu'elles sont opérationnelles et produisent des résultats tout à fait valables. La bibliothèque de calcul ne pouvant qu'évoluer positivement au fil des années, c'est un outil à suivre assurément.

# 5 Bibliographie

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. Cet article est intéressant parce qu'il propose une démarche particulièrement limpide – et reproductible par tout un chacun – pour évaluer les sorties des logiciels, en s'appuyant sur les données et les résultats fournis par le NIST (<u>Statistical Reference Datasets</u> – National Institute of Standard and Technology). On notera au passage que Google Docs est à la

traine dans le comparatif. C'est un excellent outil pour les tableaux de bords et pour le partage de documents. Il n'est pas recommandé en revanche pour les calculs statistiques (pour l'instant !).

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 ».