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 KDNuggets. Excel arrive régulièrement parmi les trois logiciels les plus utilisés ces cinq dernières années (enquêtes 2013, 2012, 2011, 2010, 2009). 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'APEC (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).

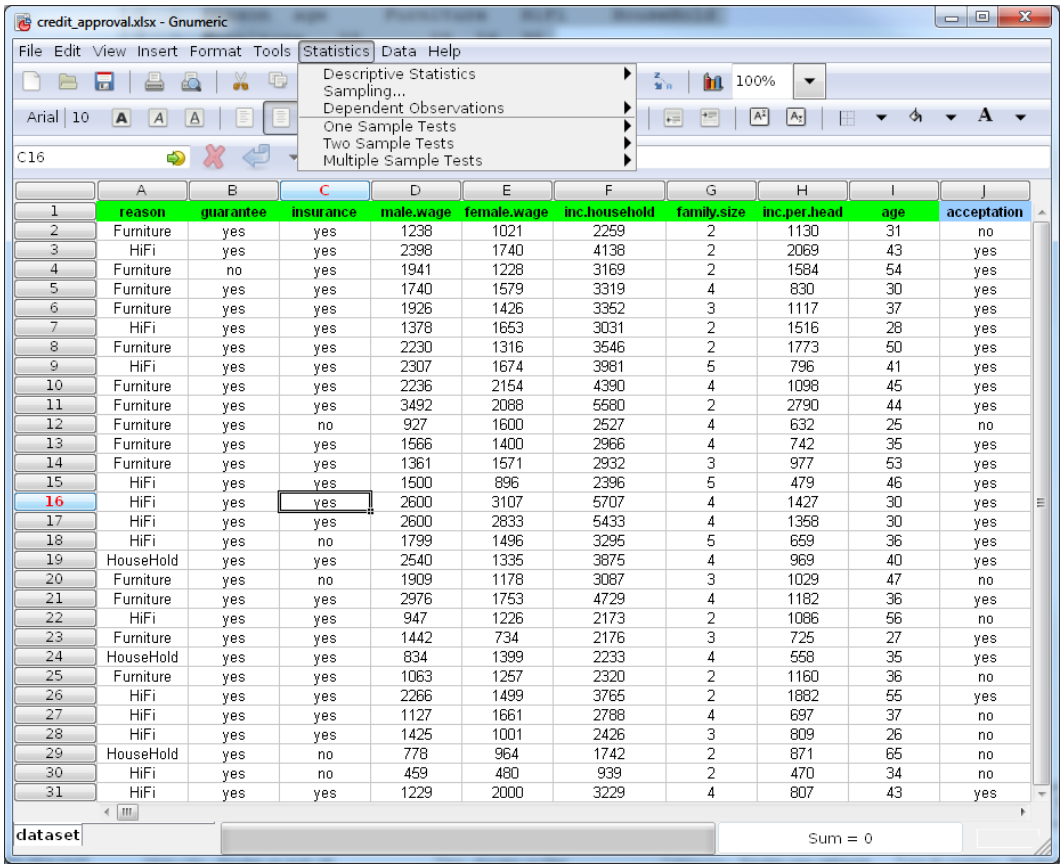
Ce succès n'est pas très étonnant à vrai dire. Le tableur est très simple à utiliser, il possède des fonctionnalités multiples, y compris celle de savoir manipuler des tableaux de données pouvant aller jusqu'à 1.048.575 observations et 16.384 variables (ça nous laisse un peu de marge). Il est très largement 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. « The Risks of Using Spreadsheets for Statistical Analysis », 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 SAS s'y est mis, idem pour Microsoft. 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** (http://www.gnumeric.org/). 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[1] ; 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.



	A	B	C	D	E	F	G	H	I	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

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

[1] 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.

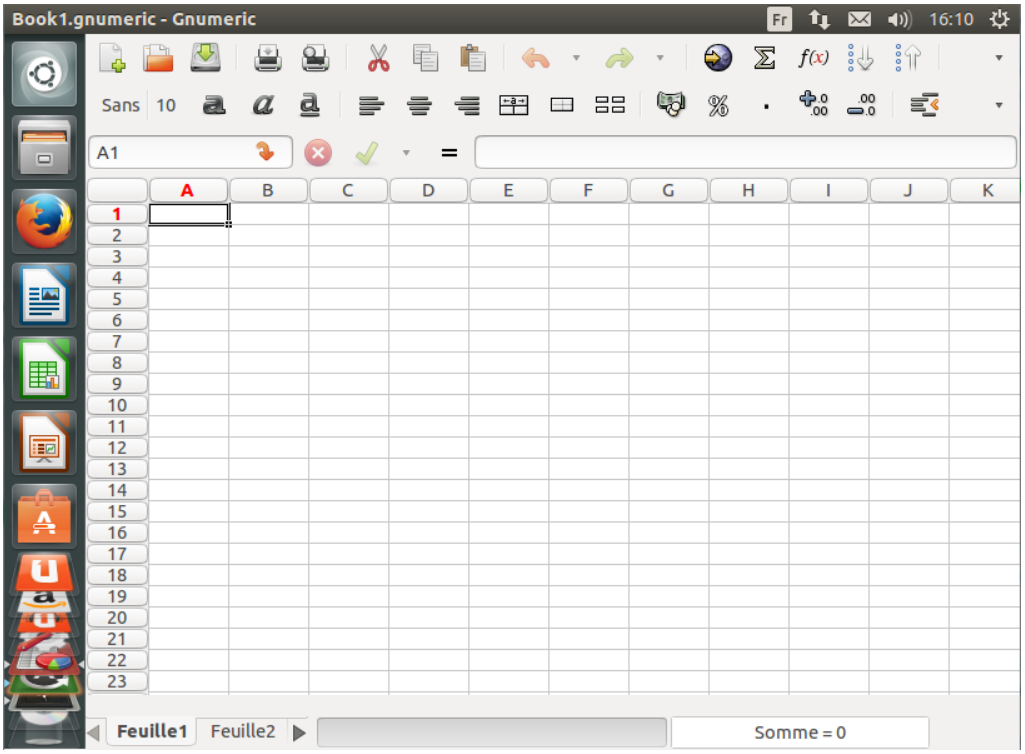


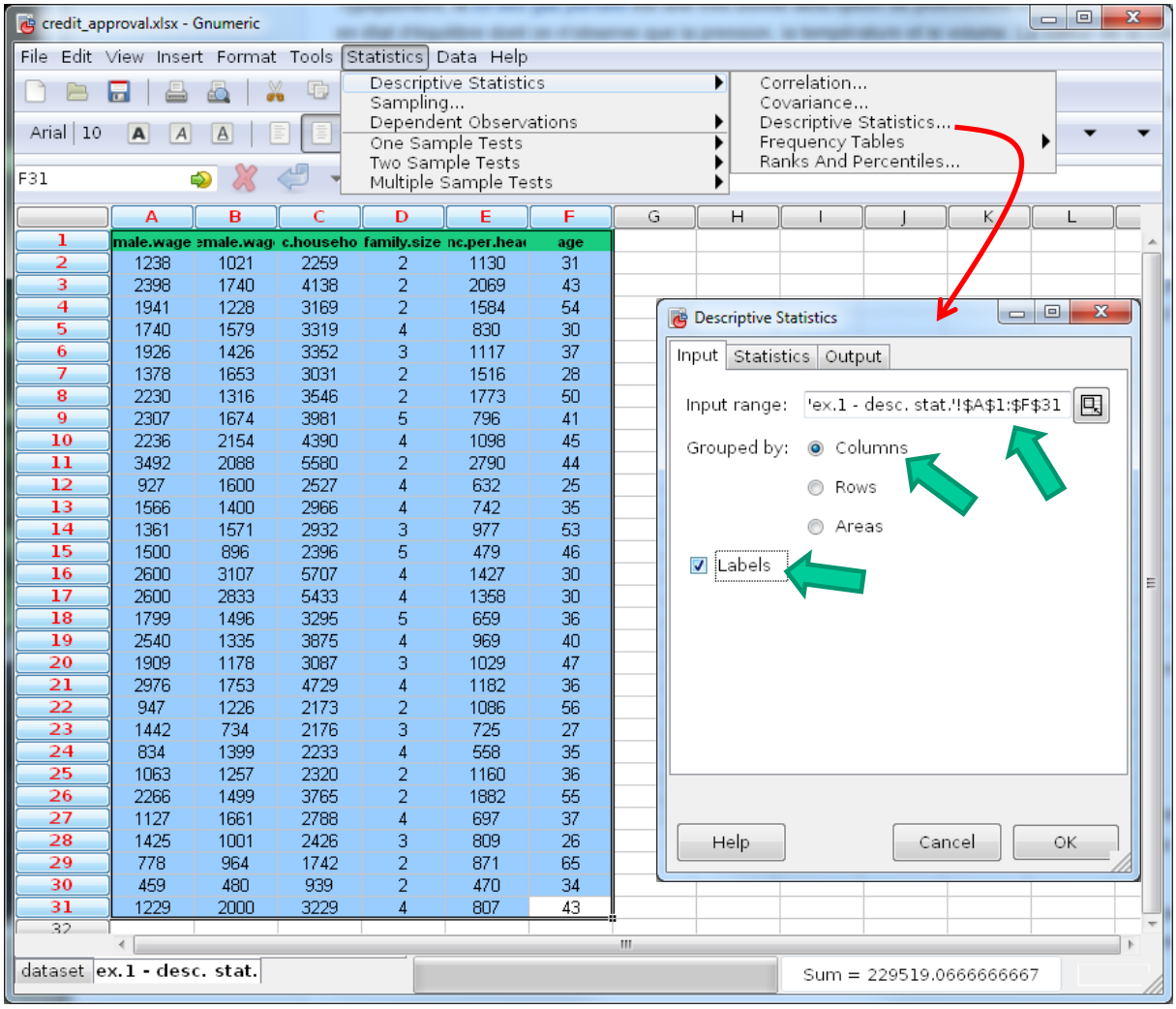
Figure 2 - Gnumeric sous Ubuntu

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

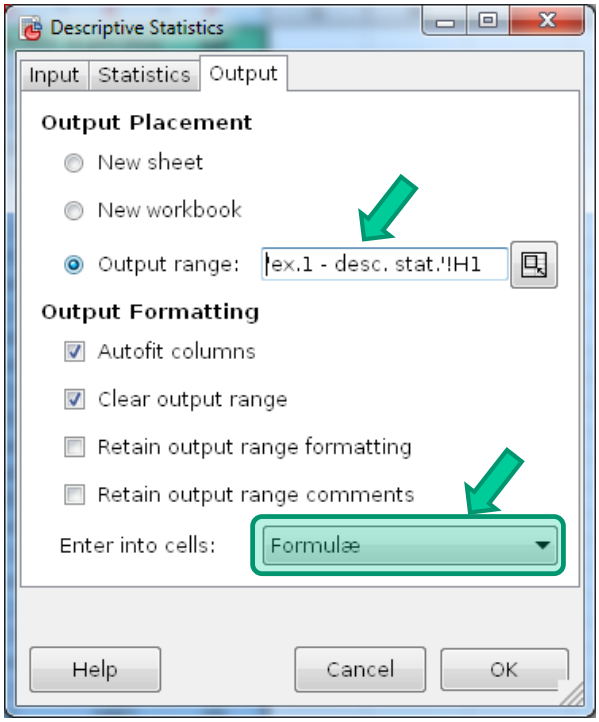
3.1 Statistiques descriptives

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 Statistics / Descriptive Statistics**. Une boîte de paramétrage apparaît :

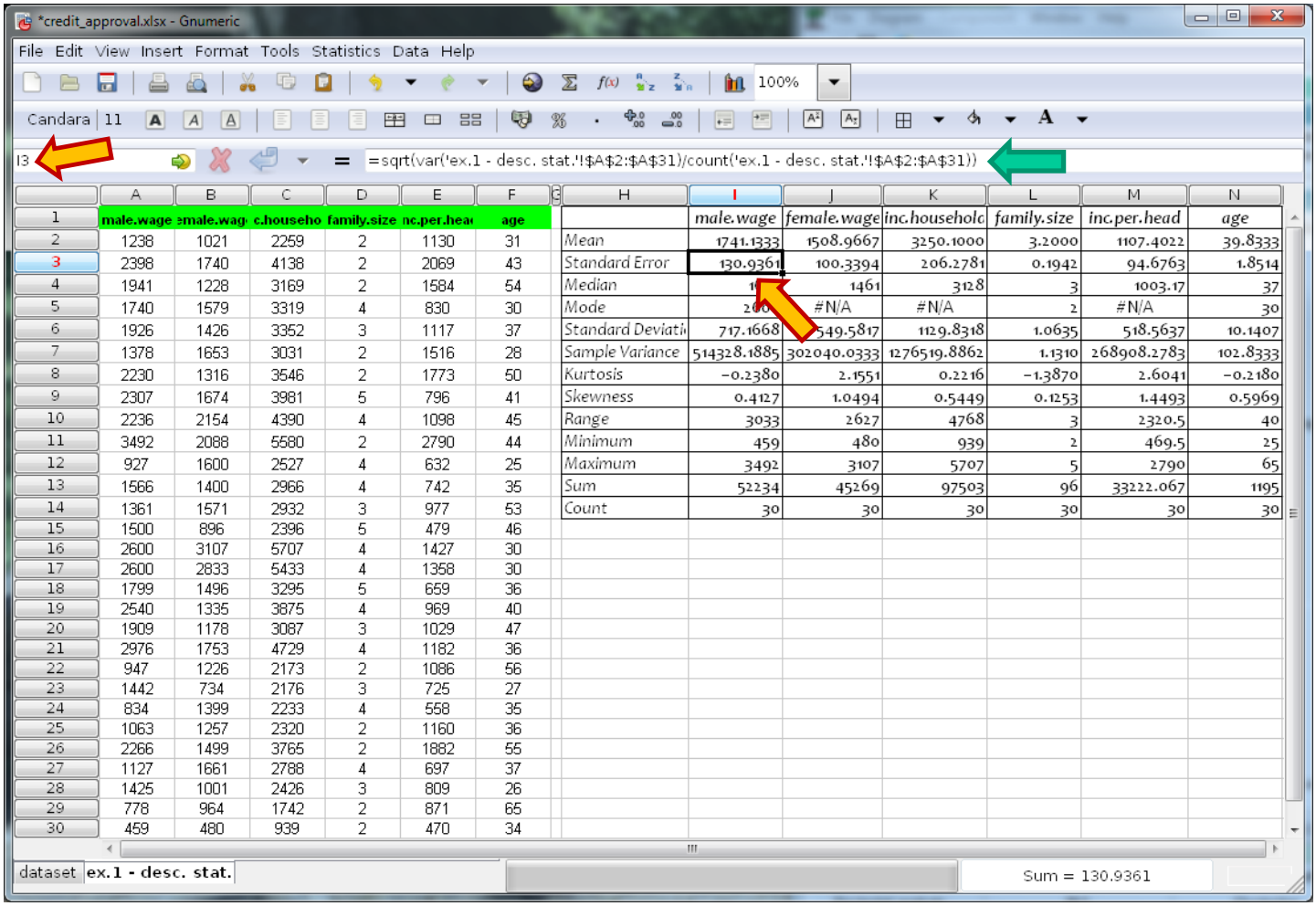


Dans l'onglet INPUT, la plage de données doit être correctement sélectionnée ; elle est organisée en colonnes ; il est très important de cocher l'option « Labels » pour signifier au logiciel que la première ligne correspond aux noms des variables.

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.



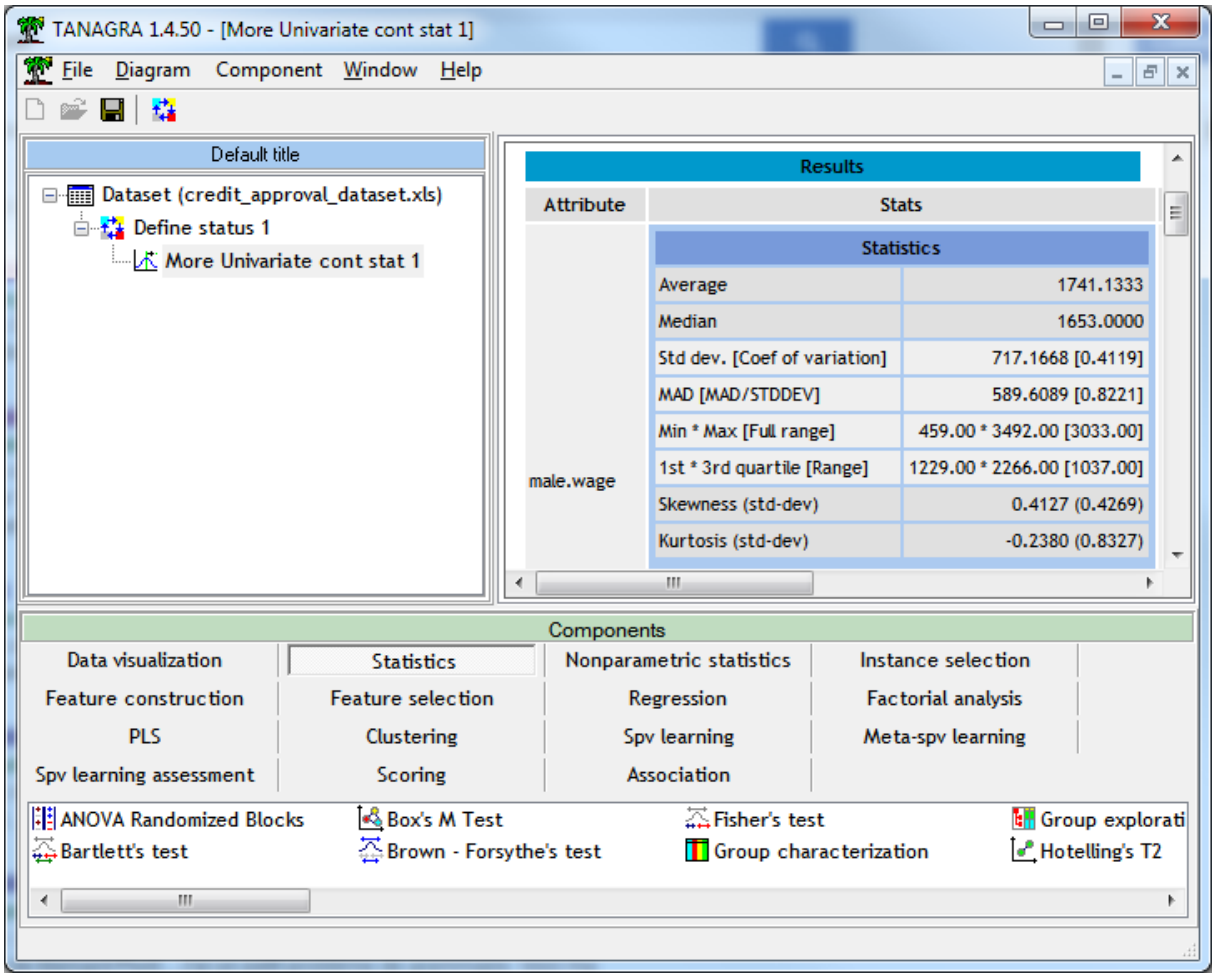
Nous obtenons, entres autres, la moyenne, la médiane, l'écart-type, etc. (les résultats ont été formatés pour rendre la lecture plus facile).



	male.wage	female.wage	inc.household	family.size	inc.per.head	age
Mean	1741.1333	1508.9667	3250.1000	3.2000	1107.4022	39.8333
Standard Error	130.9361	100.3394	206.2781	0.1942	94.6763	1.8514
Median	1	1461	3128	3	1003.17	37
Mode	200	#N/A	#N/A	2	#N/A	30
Standard Deviati	717.1668	549.5817	1129.8318	1.0635	518.5637	10.1407
Sample Variance	514328.1885	302040.0333	1276519.8862	1.1310	268908.2783	102.8333
Kurtosis	-0.2380	2.1551	0.2216	-1.3870	2.6041	-0.2180
Skewness	0.4127	1.0494	0.5449	0.1253	1.4493	0.5969
Range	3033	2627	4768	3	2320.5	40
Minimum	459	480	939	2	469.5	25
Maximum	3492	3107	5707	5	2790	65
Sum	52234	45269	97503	96	33222.067	1195
Count	30	30	30	30	30	30

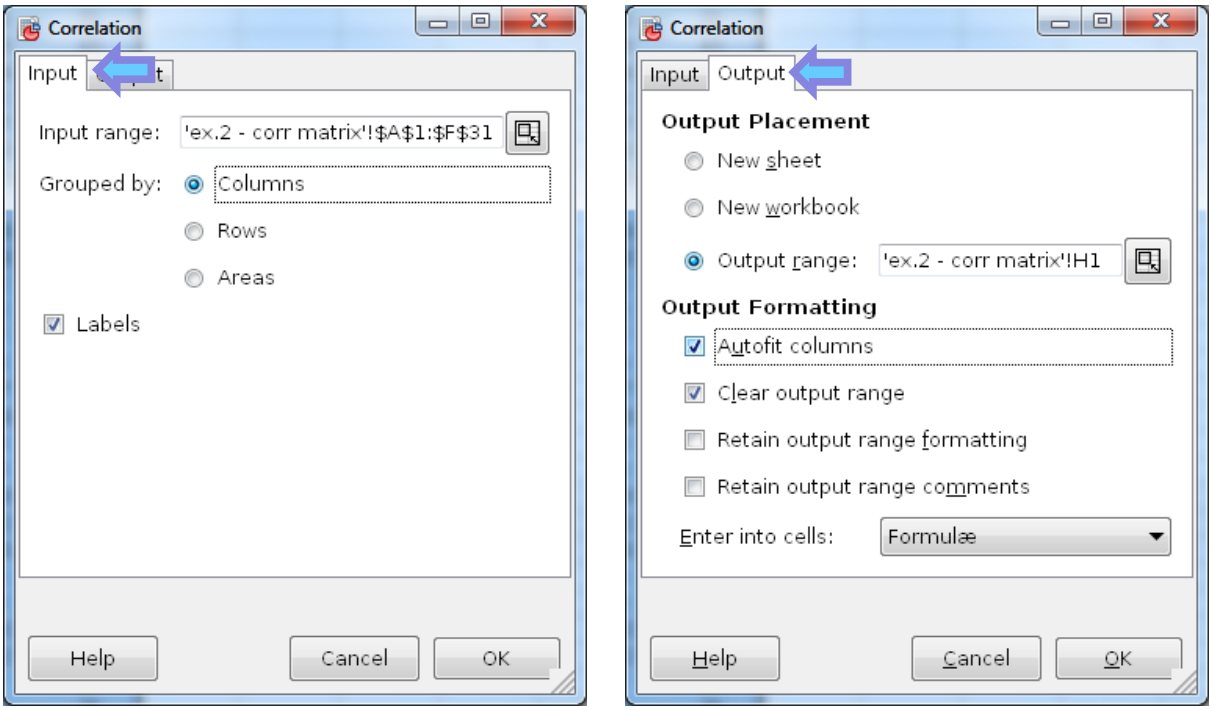
Attardons-nous un instant sur l'écart-type de la moyenne de « X : male.wage ». En cellule I3, nous distinguons la formule $s_{\bar{x}} = \sqrt{\frac{s_x^2}{n}} = \sqrt{\frac{514328.1885}{30}} = 130.9361$. La variance estimée s_x^2 de X est en cellule I7.

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

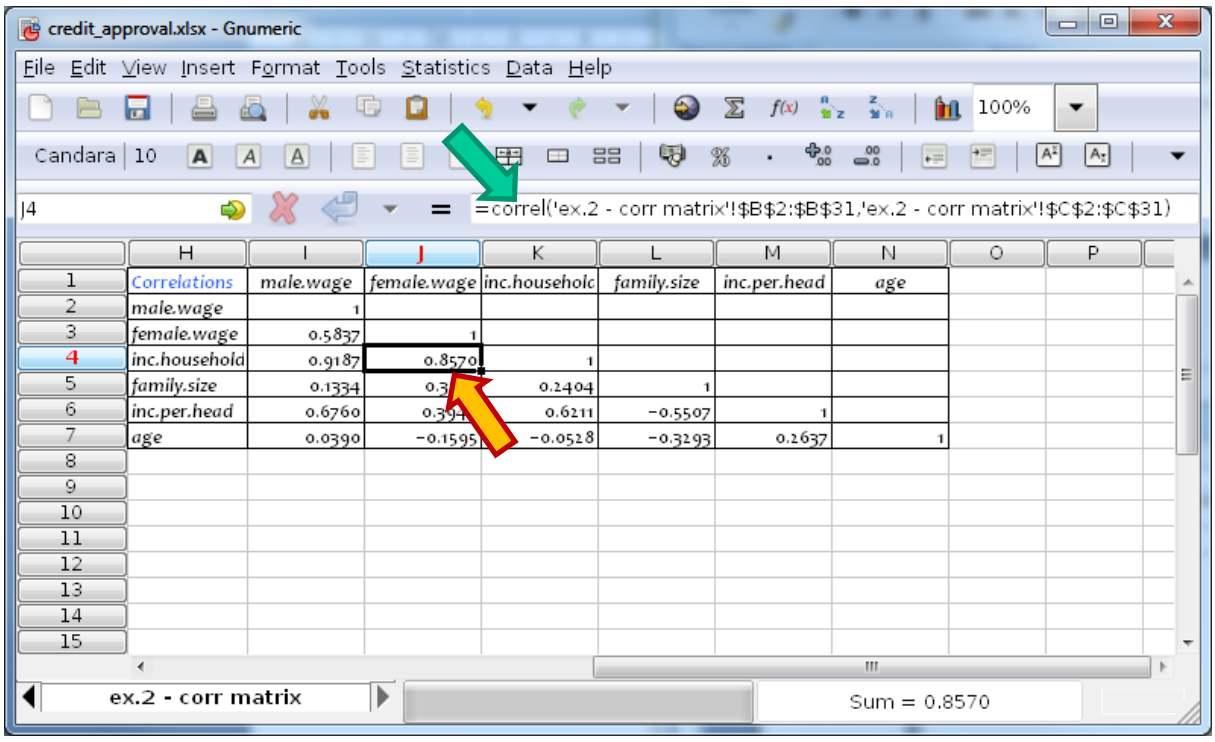


3.2 Matrice des corrélations

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.



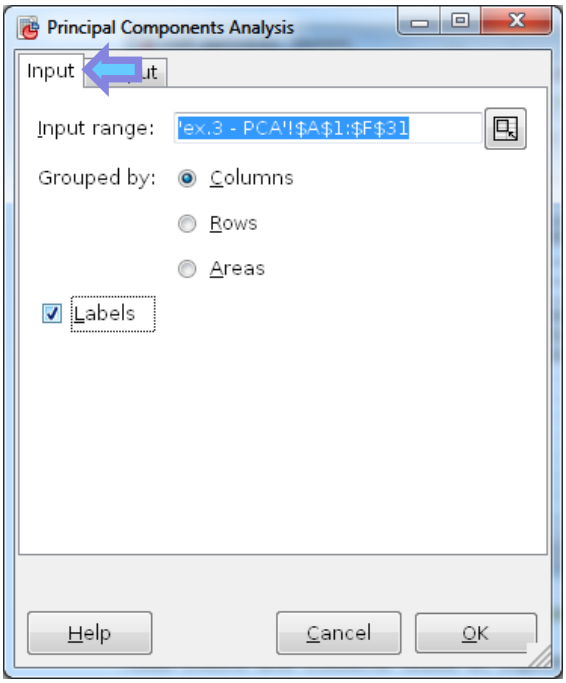
Les corrélations sont obtenues à partir de la formule CORREL.

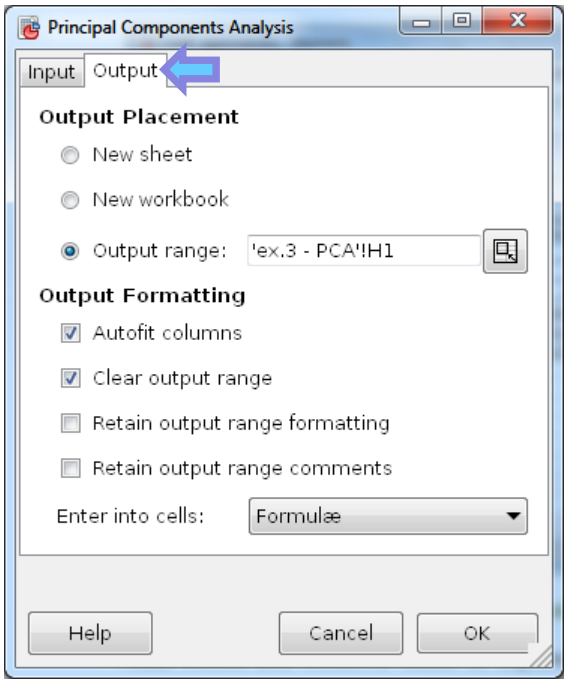


	H	I	J	K	L	M	N
1	Correlations	male.wage	female.wage	inc.household	family.size	inc.per.head	age
2	male.wage	1					
3	female.wage	0.5837	1				
4	inc.household	0.9187	0.8570	1			
5	family.size	0.1334	0.3	0.2404	1		
6	inc.per.head	0.6760	0.394	0.6211	-0.5507	1	
7	age	0.0390	-0.1595	-0.0528	-0.3293	0.2637	1

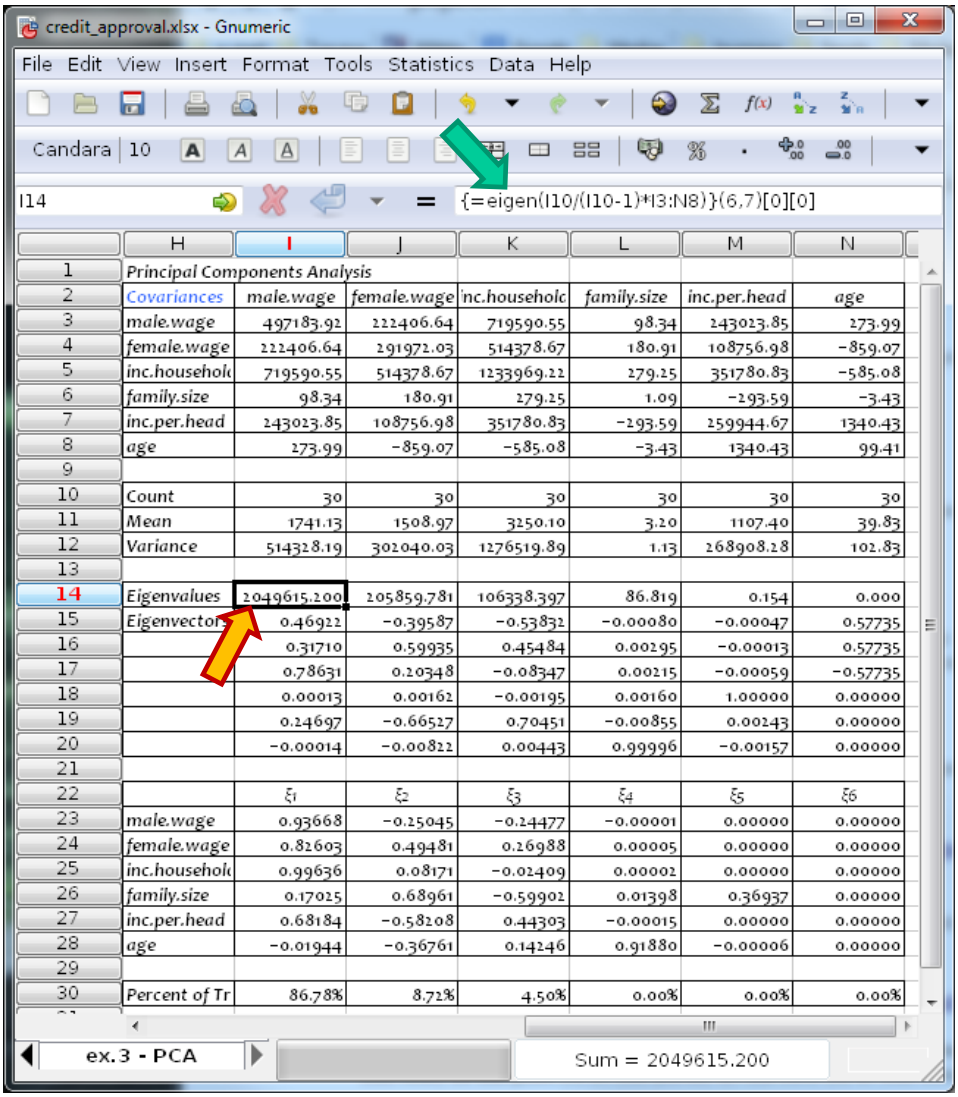
3.3 Analyse en composantes principales

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 :





Nous obtenons :



	H	I	J	K	L	M	N
1	*Principal Components Analysis*						
2	*Covariances*	*male.wage*	*female.wage*	*inc.household*	*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
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_1	ξ_2	ξ_3	ξ_4	ξ_5	ξ_6
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%

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) :

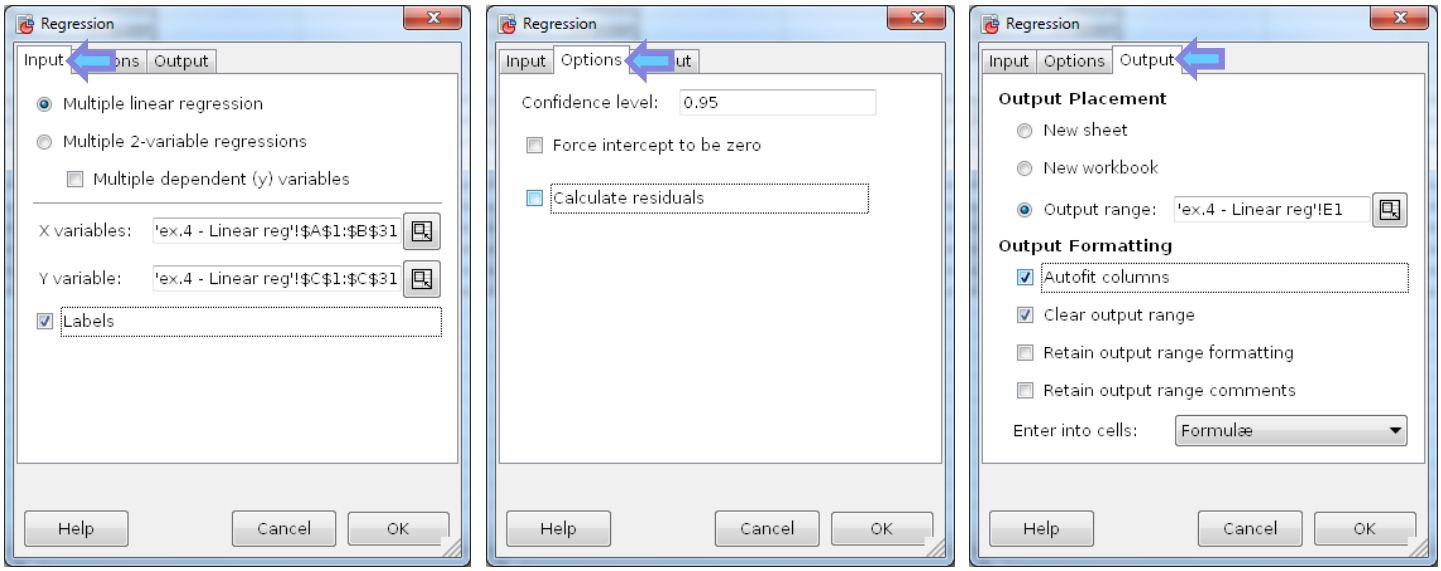
Axis	Eigen value	Difference	Proportion (%)	Cumulative (%)
1	1981294.694	1782296.905	86.78%	86.78%
2	198997.789	96204.006	8.72%	95.49%
3	102793.782	102709.858	4.50%	100.00%
4	83.925	83.776	0.00%	100.00%
5	0.149	0.149	0.00%	100.00%
6	0	-	0.00%	100.00%
Tot.	2283170.339	-	-	-

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, λ_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.

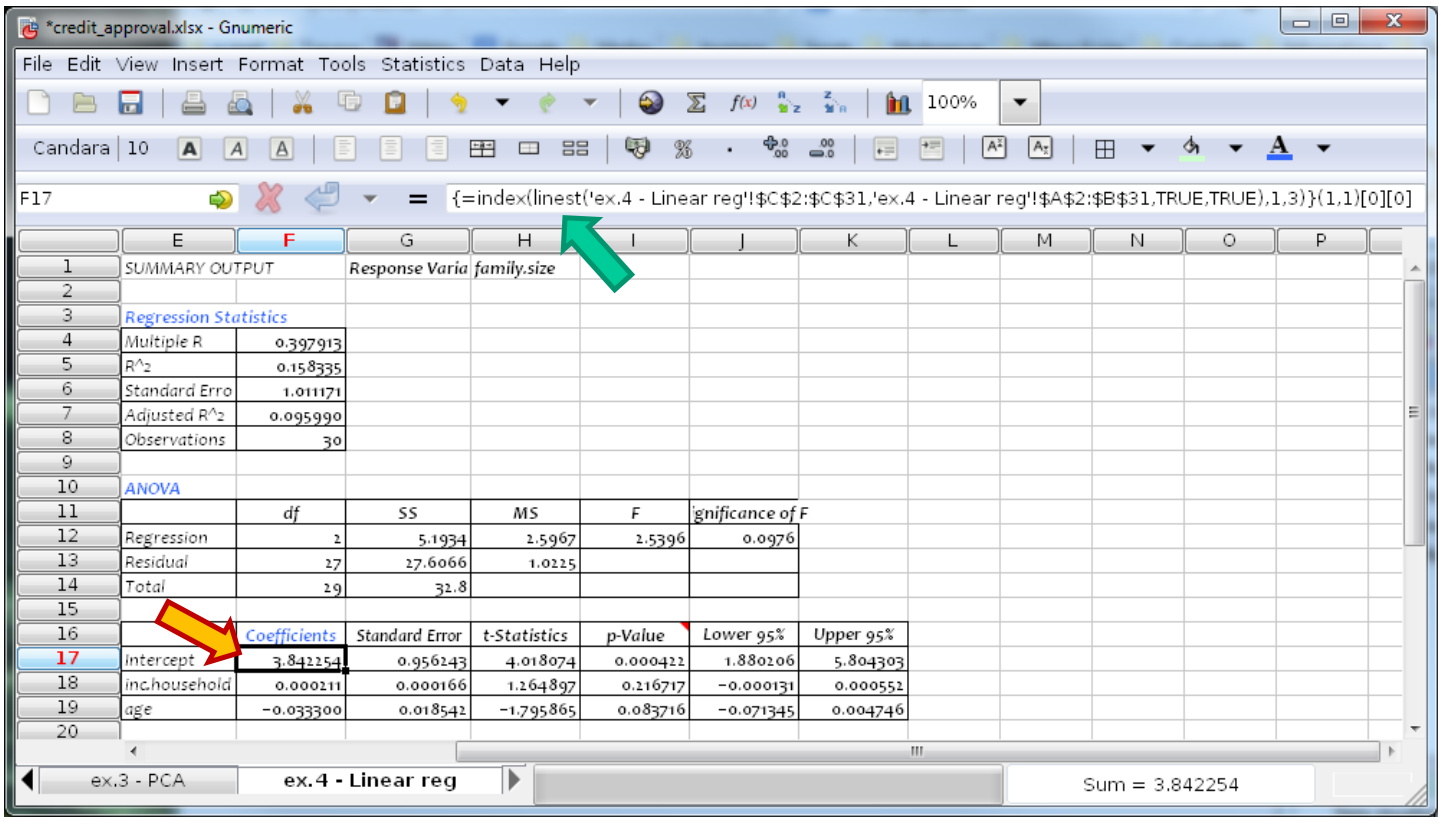
3.4 Régression linéaire

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.

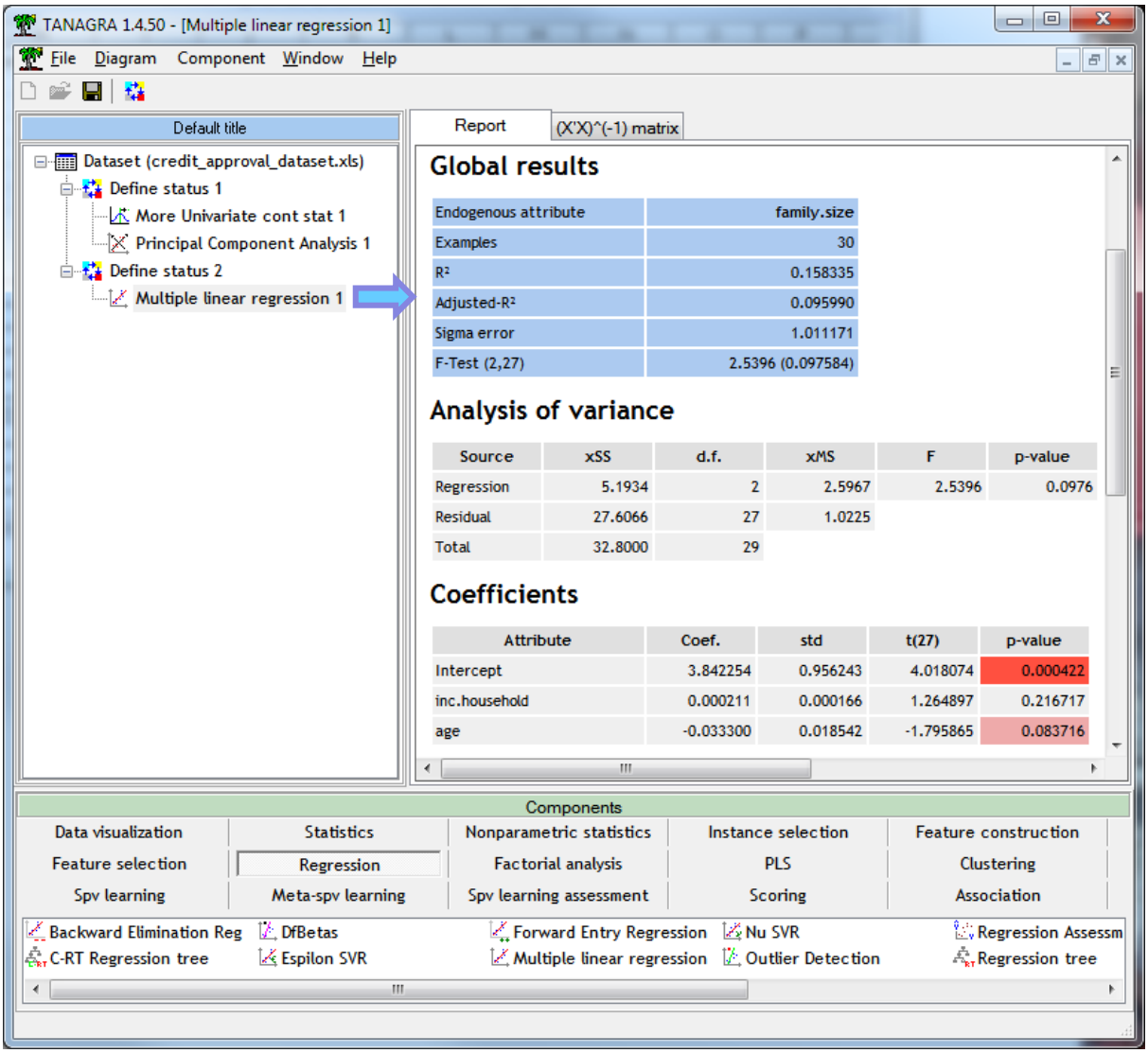


Y est la variable dépendante (familiy.size), X représente le bloc des variables indépedantes (inc.household et age). Gnumeric s'appuie essentiellement sur la fonction LINEST. Il réorganise les résultats pour une présentation plus conforme aux sorties des logiciels de statistique. Il pioche les différentes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



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).

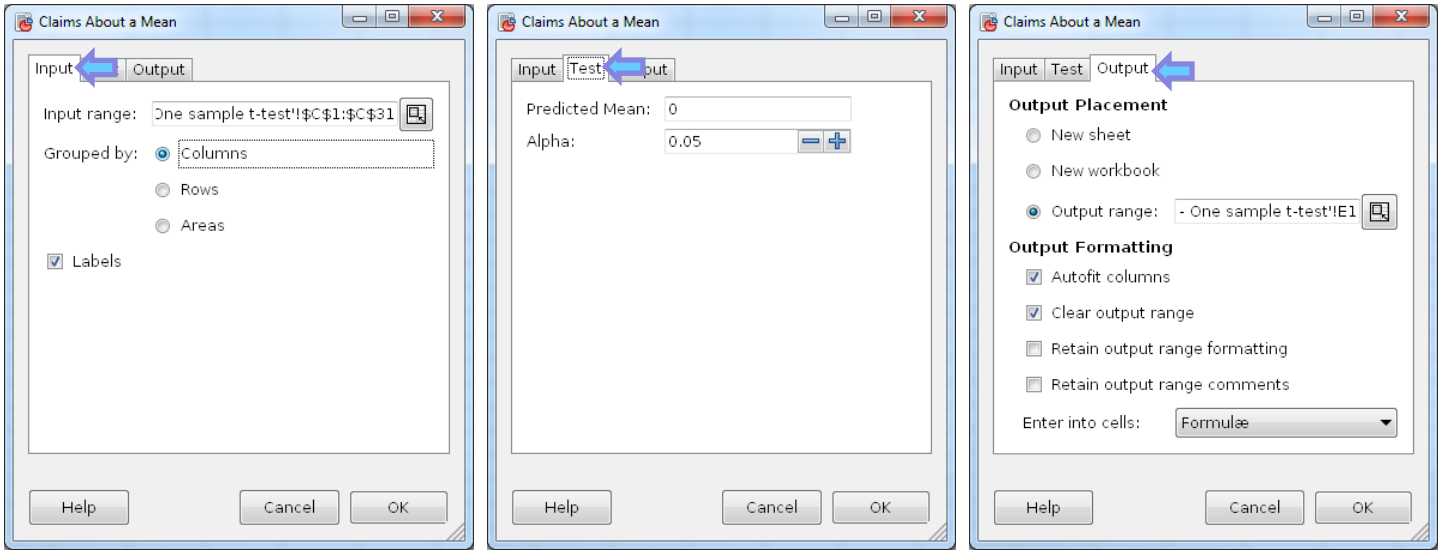
Nous retrouvons les mêmes résultats avec Tanagra, l'organisation est identique.



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.

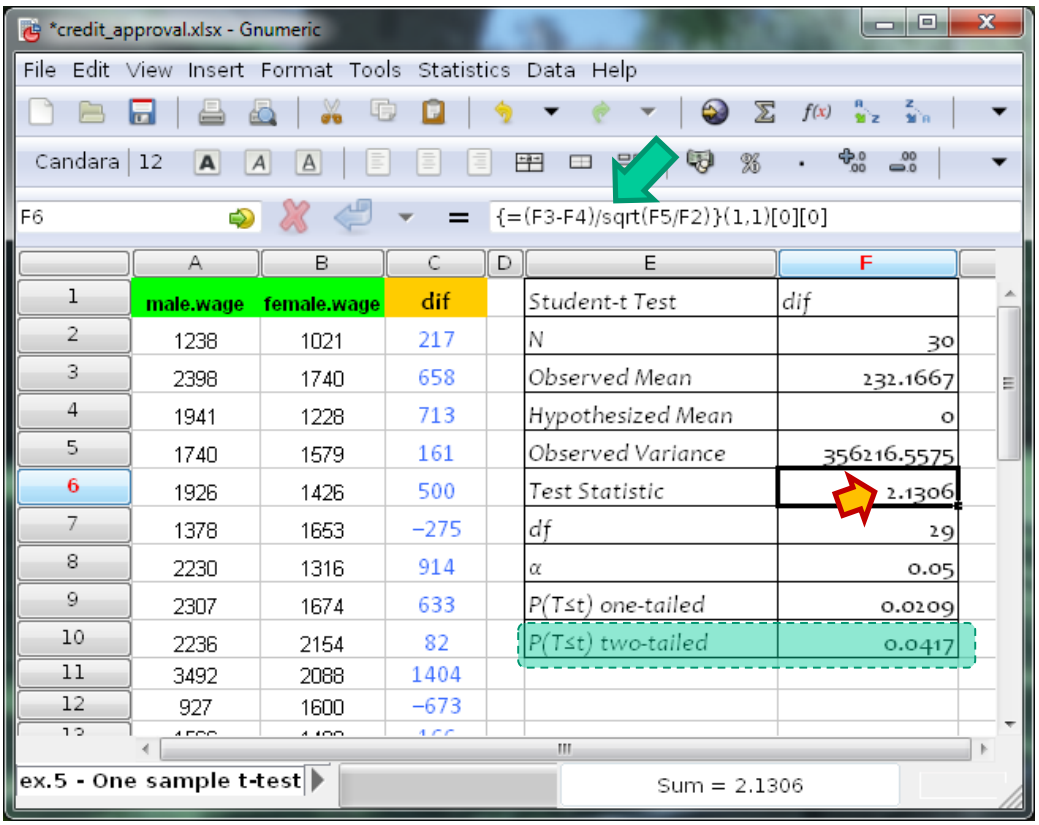
Après avoir sélectionné la colonne DIF, nous actionnons le menu **Statistics / One Sample Tests / Claims About a Mean**. Nous paramétrons la procédure comme suit.



Au risque 5%, nous rejetons l'hypothèse de la nullité (Predicted Mean = 0 dans l'onglet TEST) de la moyenne de DIF[2]. Nous observons dans la cellule F6 la formule (μ_0 = 0 dans notre exemple) :

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

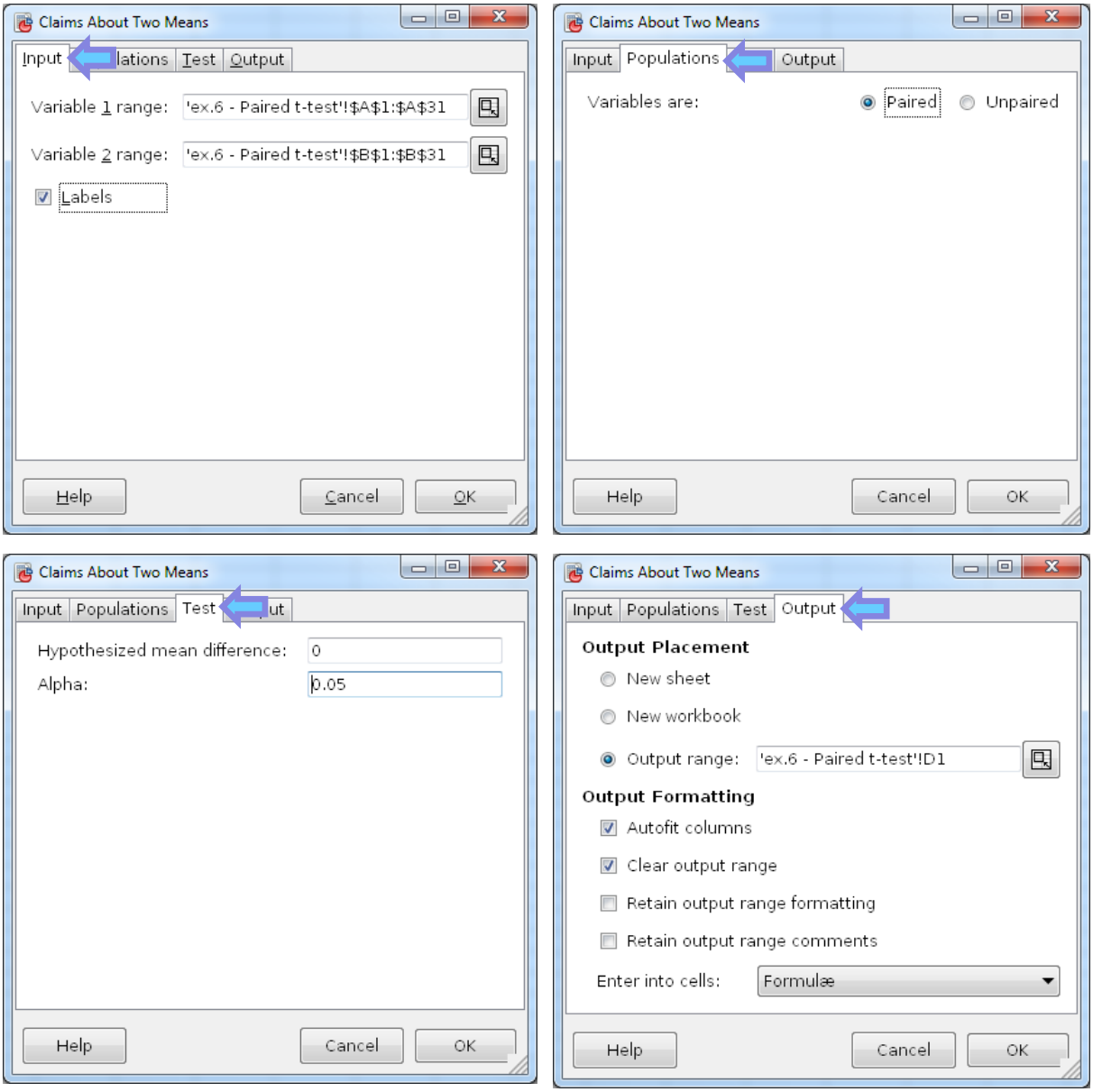
La p-value est obtenue à l'aide de la fonction TDIST.



[2] 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**].

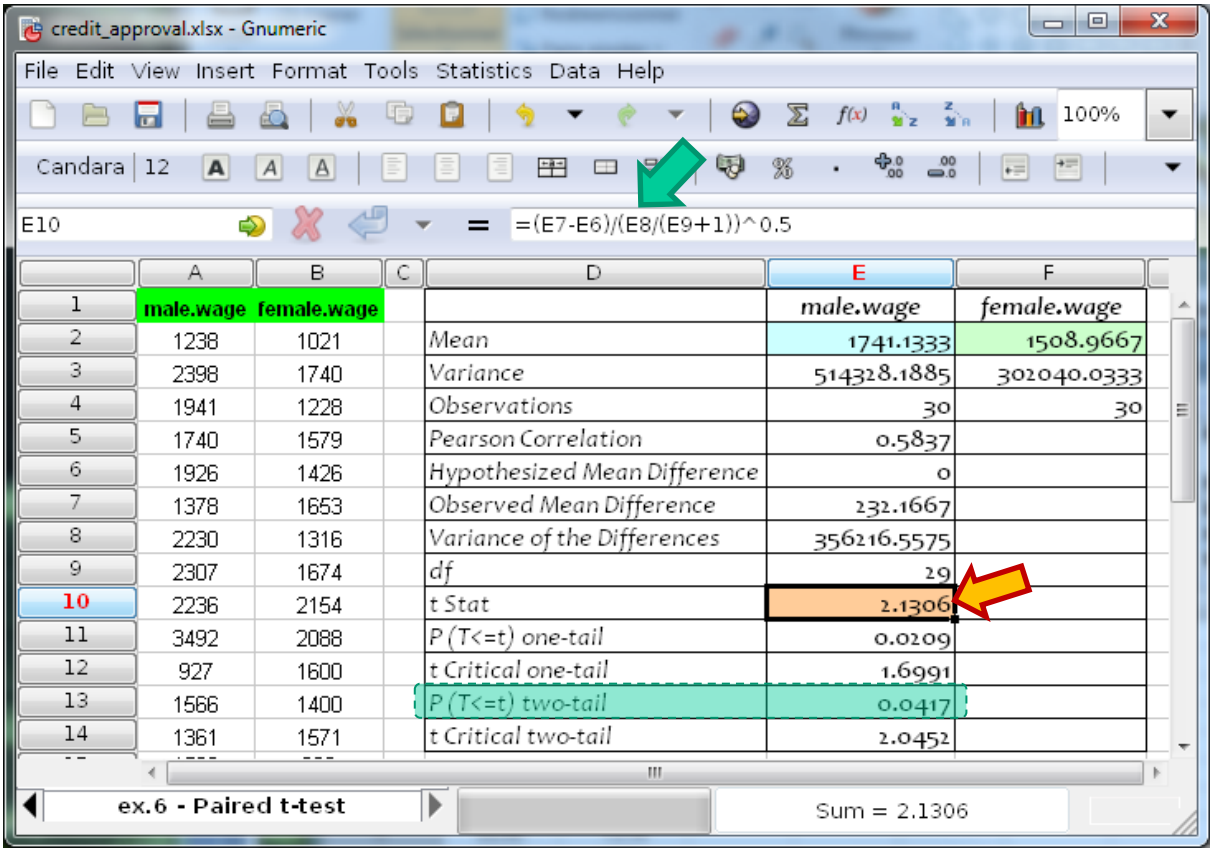
3.6 Comparaisons de moyennes – Echantillons appariés

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[3]. 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.

[3] R. Rakotomalala, « Comparaison de populations - Tests paramétriques », version 1.2, Juin 2013 ; http://eric.univ-lyon2.fr/~ricco/cours/cours/Comp_Pop_Tests_Parametriques.pdf



	A	B	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
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	

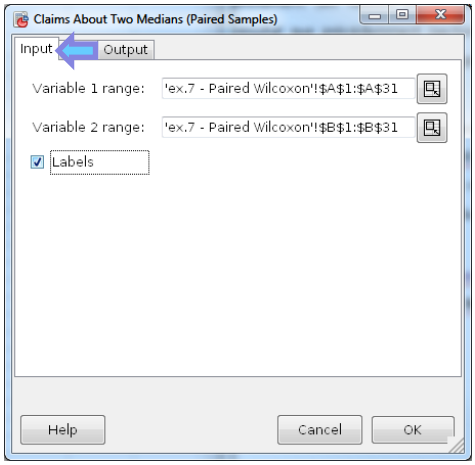
Par un procédé différent, nous obtenons strictement (les valeurs et distributions des statistiques de test sont identiques) le même résultat que précédemment (section 3.5). Les hommes et les femmes ont des niveaux de salaire différents à l'intérieur des ménages.

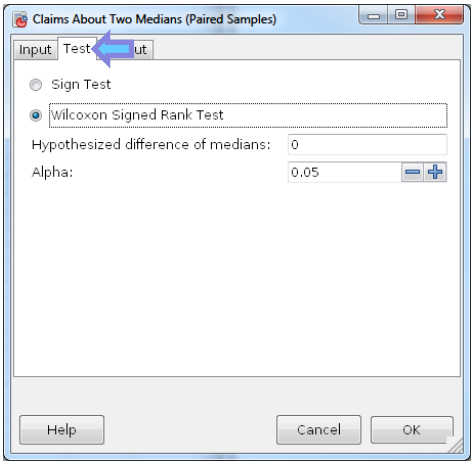
3.7 Test non-paramétrique – Echantillons appariés

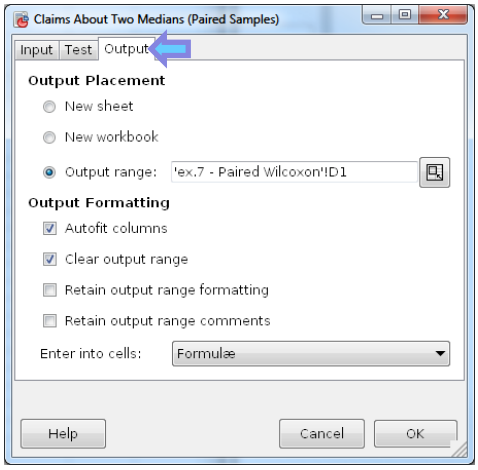
Nous pouvons répondre à la même question, comparer les salaires des hommes et des femmes à l'intérieur des ménages, en utilisant le test des rangs signés de Wilcoxon. La statistique de test est basée sur le rang des écarts, et non sur leur amplitude[4]. L'intérêt de cette procédure est que l'on s'affranchit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

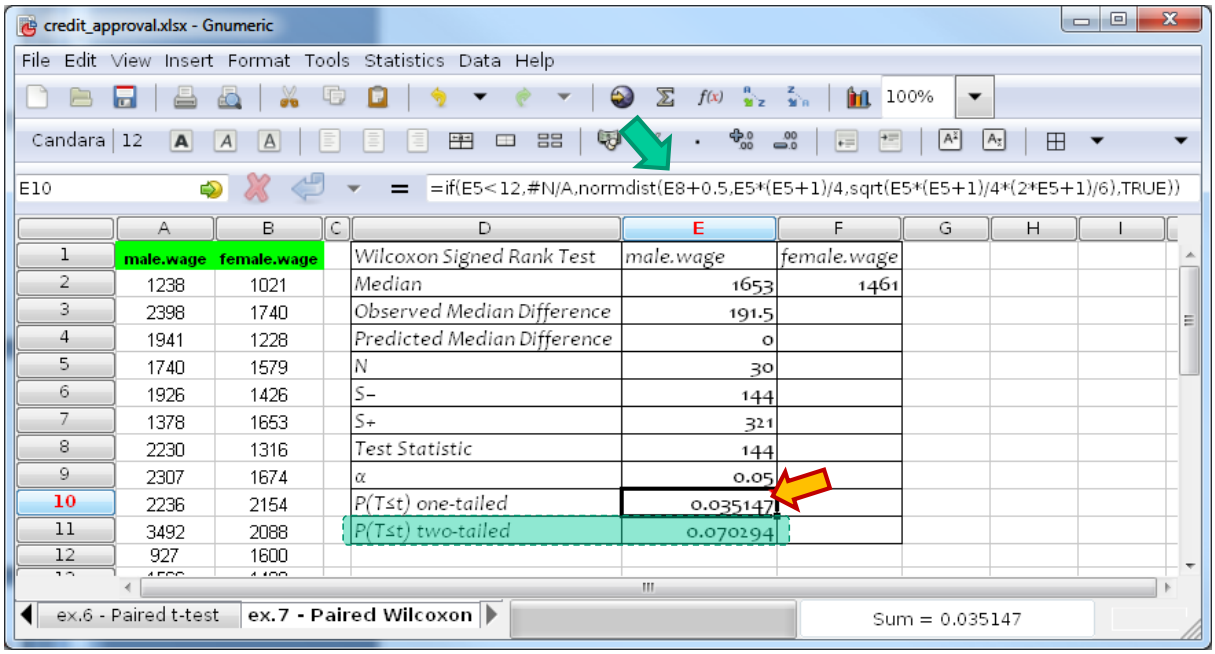
[4] R. Rakotomalala, « Comparaisons de populations – Tests non paramétriques », Août 2008 ; http://eric.univ-lyon2.fr/~ricco/cours/cours/Comp_Pop_Tests_Nonparametriques.pdf (pages 130 et suivantes).







Nous obtenons :



Les effectifs étant suffisamment élevés ($n \geq 12$), 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		Attribute_X		Statistical test	
male.wage		female.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

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 :

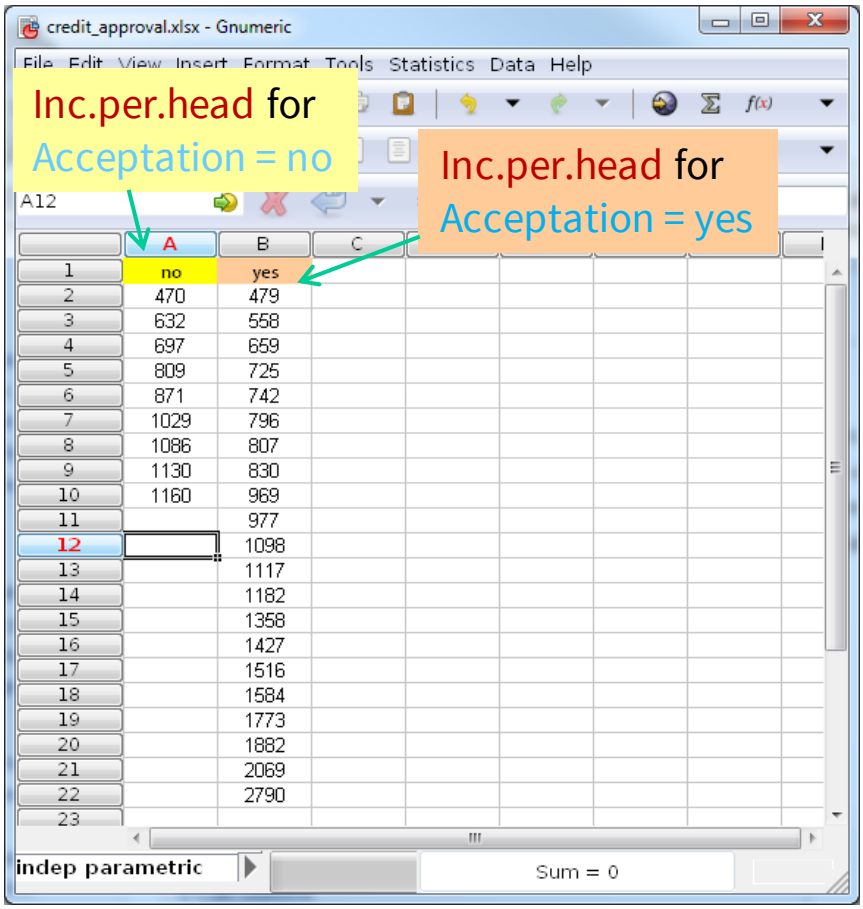
$$p.value = 2\times[1 - \Phi(1.810014)] = 0.070294$$

Exactement la valeur produite par Gnumeric.

3.8 Test paramétrique – Echantillons indépendants

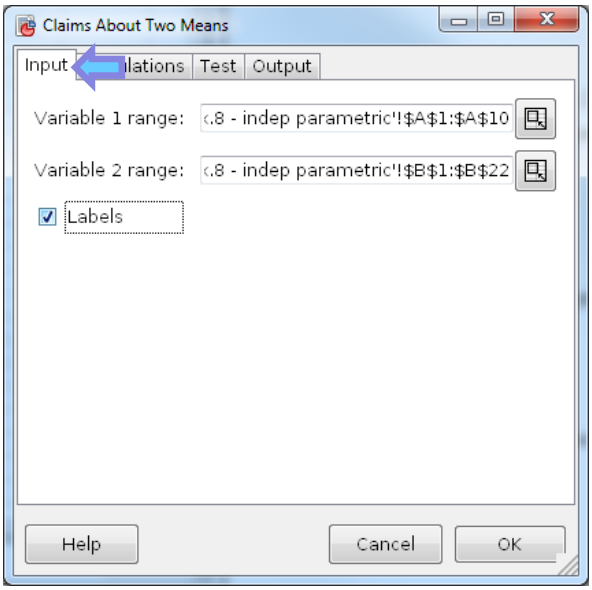
Notre propos est de comparer les revenus par tête des ménages selon l'acceptation de la demande de crédit. Il s'agit d'un test pour échantillons indépendants puisque la variable « acceptation » induit une partition non recouvrante des individus.

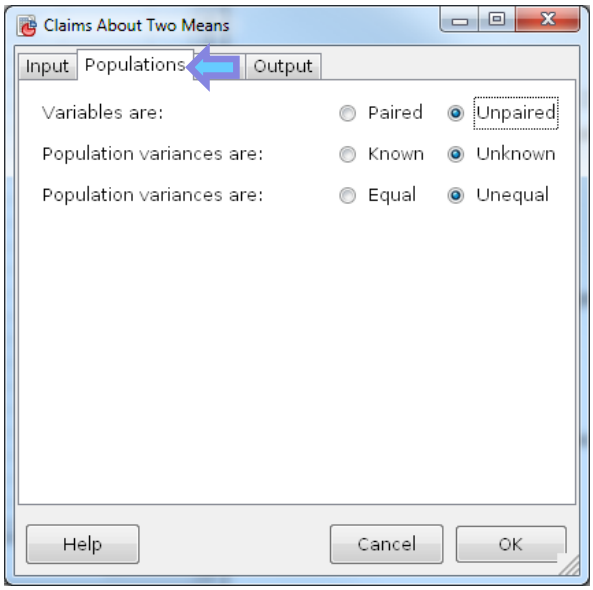
Ce test nécessite une mise en forme particulière des données dans Gnumeric dans la feuille « **ex.8 – indep parametric** ». Plutôt que la présentation initiale individus x variables, nous devons créer des colonnes de valeurs « inc.per.head » pour chaque modalité de « acceptation » (yes, no). Ces 2 colonnes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

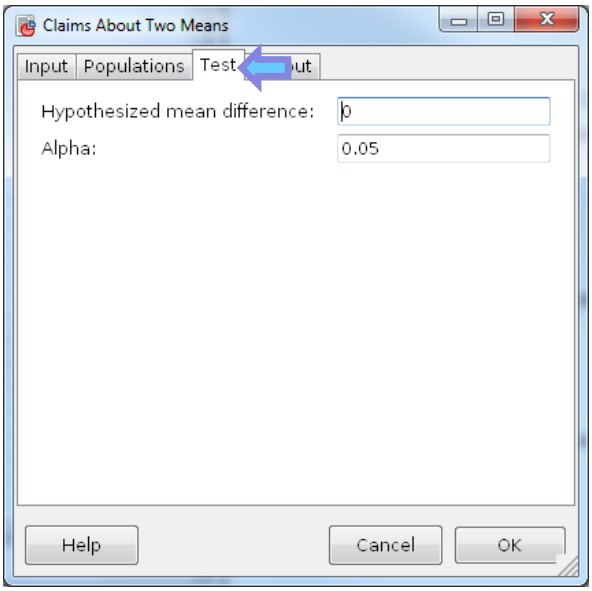


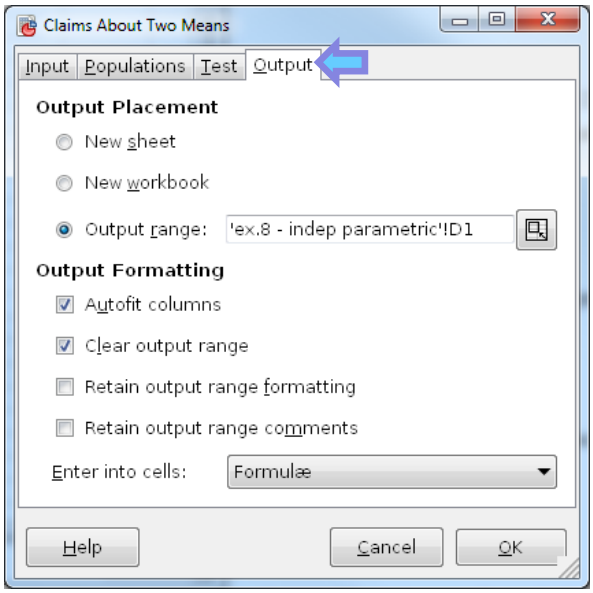
	A	B
1	no	yes
2	470	479
3	632	558
4	697	659
5	809	725
6	871	742
7	1029	796
8	1086	807
9	1130	830
10	1160	969
11		977
12		1098
13		1117
14		1182
15		1358
16		1427
17		1516
18		1584
19		1773
20		1882
21		2069
22		2790
23		

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.



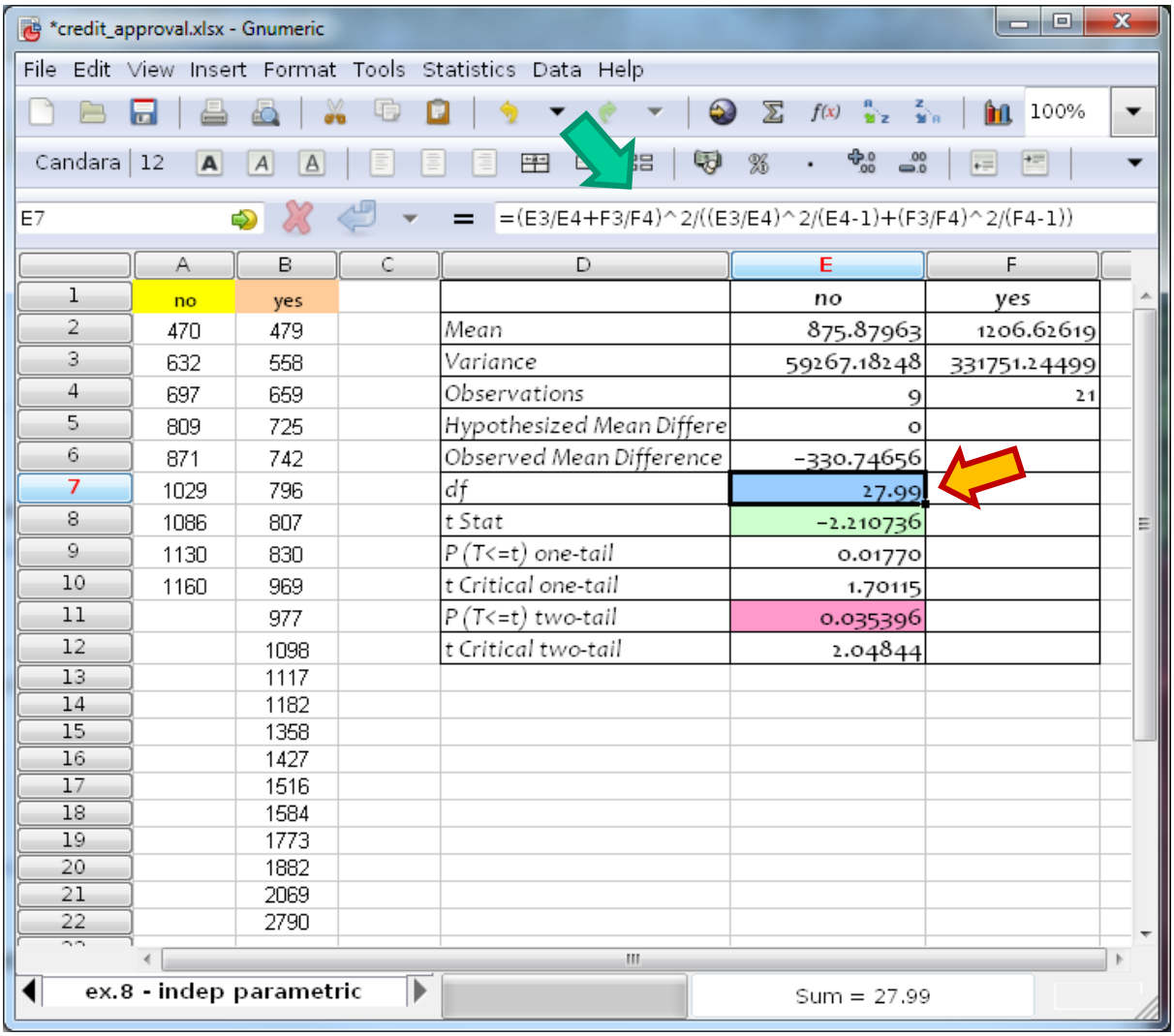






La procédure s'appuie sur le test de Welch (http://en.wikipedia.org/wiki/Welch's_t_test). La statistique est relativement facile à calculer, elle suit une loi de Student sous l'hypothèse nulle (les moyennes sont identiques). Le véritable enjeu est dans le calcul des degrés de liberté.

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.



	A	B	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	0	
6	871	742		Observed Mean Difference	-330.74656	
7	1029	796		df	27.99	
8	1086	807		t Stat	-2.210736	
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				
17		1516				
18		1584				
19		1773				
20		1882				
21		2069				
22		2790				

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

Attribute_Y	Attribute_X	Description				Statistical test	
		Value	Examples	Average	Std-dev	T	-330.7466 / 149.6092 = -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		

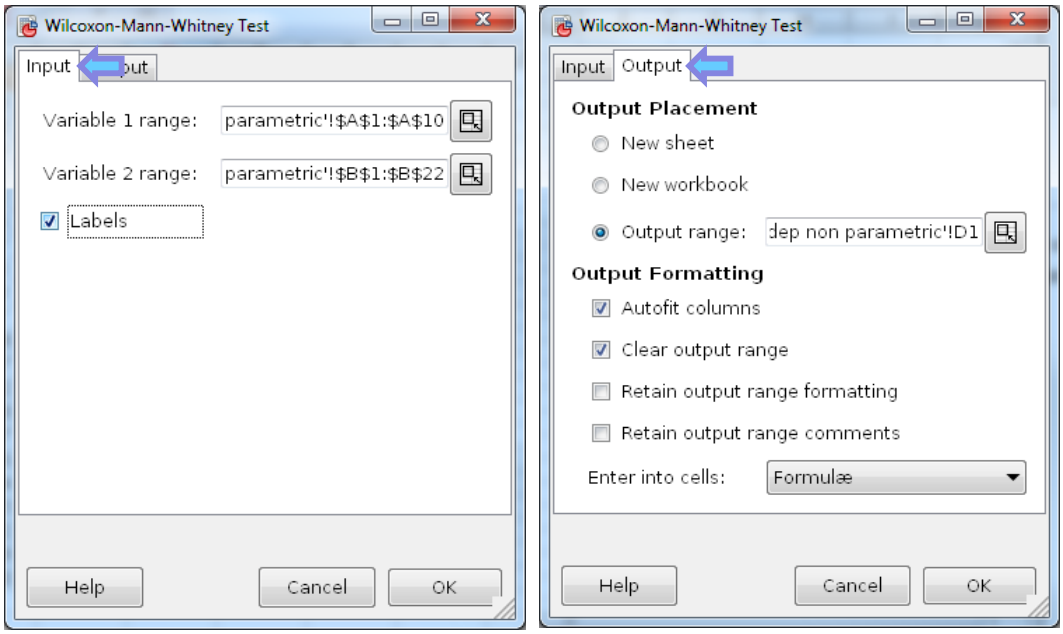
Cette différence s'explique par la gestion des degrés de liberté fractionnaires : Tanagra utilise sur l'entier le plus proche (df = 28 dans notre cas) ; la fonction TDIST de Gnumeric semble s'appuyer sur une interpolation linéaire[5] (df = entre 27 et 28). L'écart est très minime quoiqu'il en soit. Il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

3.9 Test non paramétrique – Echantillons indépendants

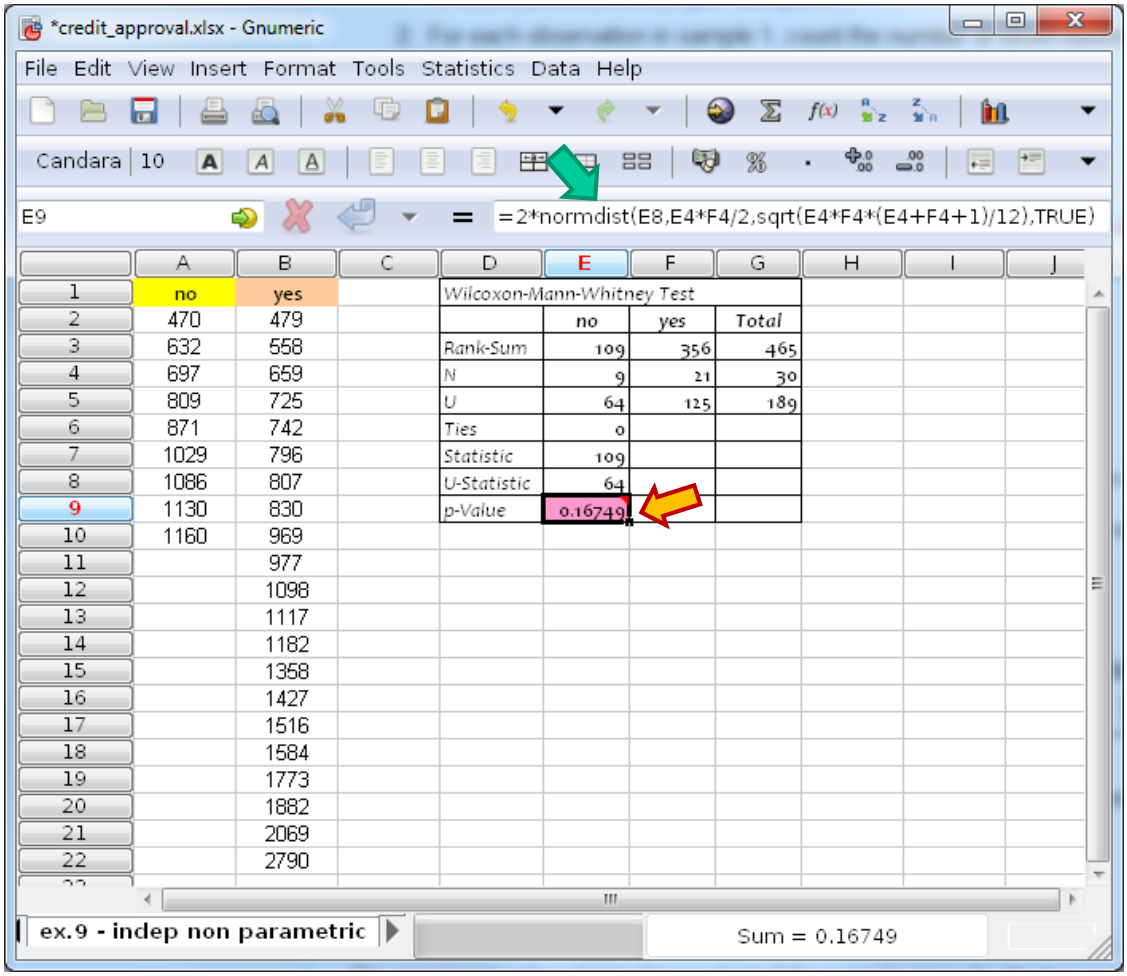
Nous utilisons le test de Wilcoxon-Mann-Whitney pour le versant non paramétrique. Les données (feuille « **ex. 9 indep non parametric** ») doivent être organisés comme

[5] 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.



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.



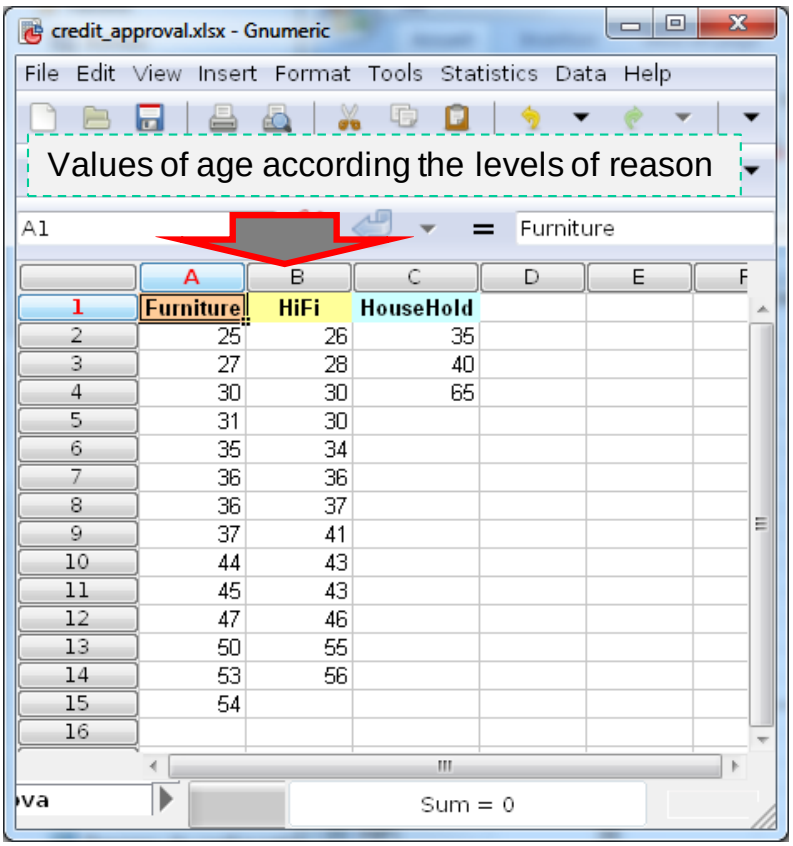
Tanagra propose exactement le même résultat, mais via une présentation différente.

		Value	Examples	Average	Rank sum	Rank mean	Mann-Whitney U	64
inc.per.head	acceptation	no	9	875.8796	109	12.1111	E(U)	94.5
		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

3.10 Analyse de variance (ANOVA)

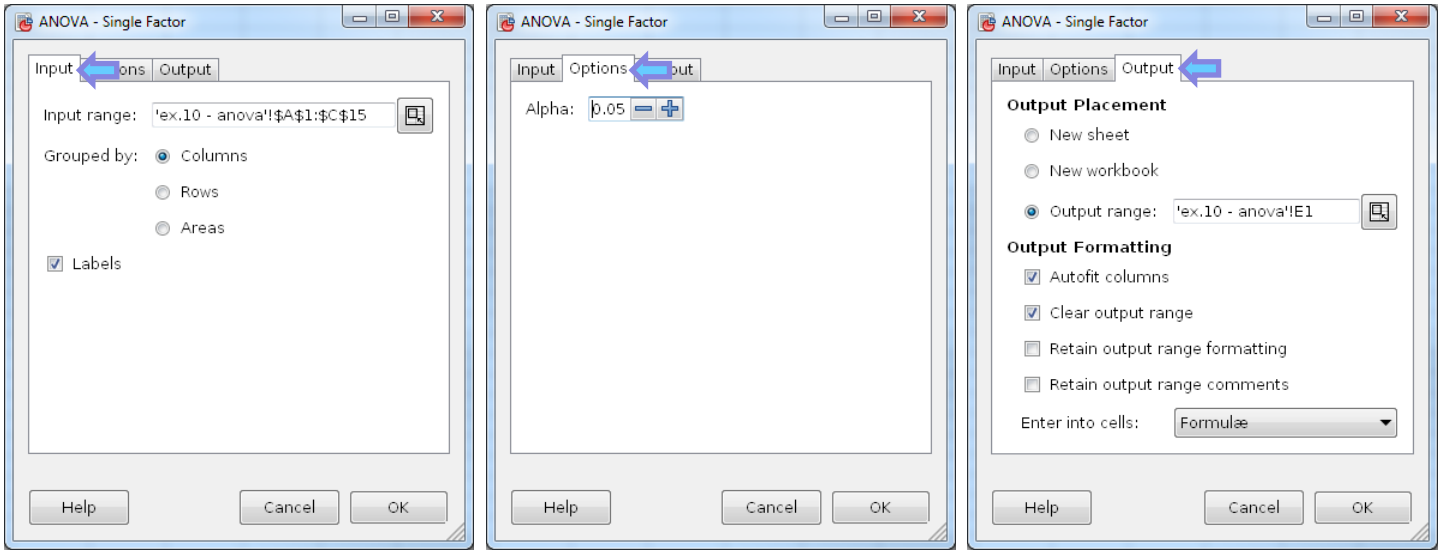
Nous souhaitons maintenant comparer l'âge des personnes selon le motif (reason) de la demande de crédit. Nous créons la feuille « **ex.10 – anova** ». Nous créons autant de listes de valeurs (âge) qu'il y a de modalités de « reason » (furniture, hifi et household). Les effectifs conditionnels ne sont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

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

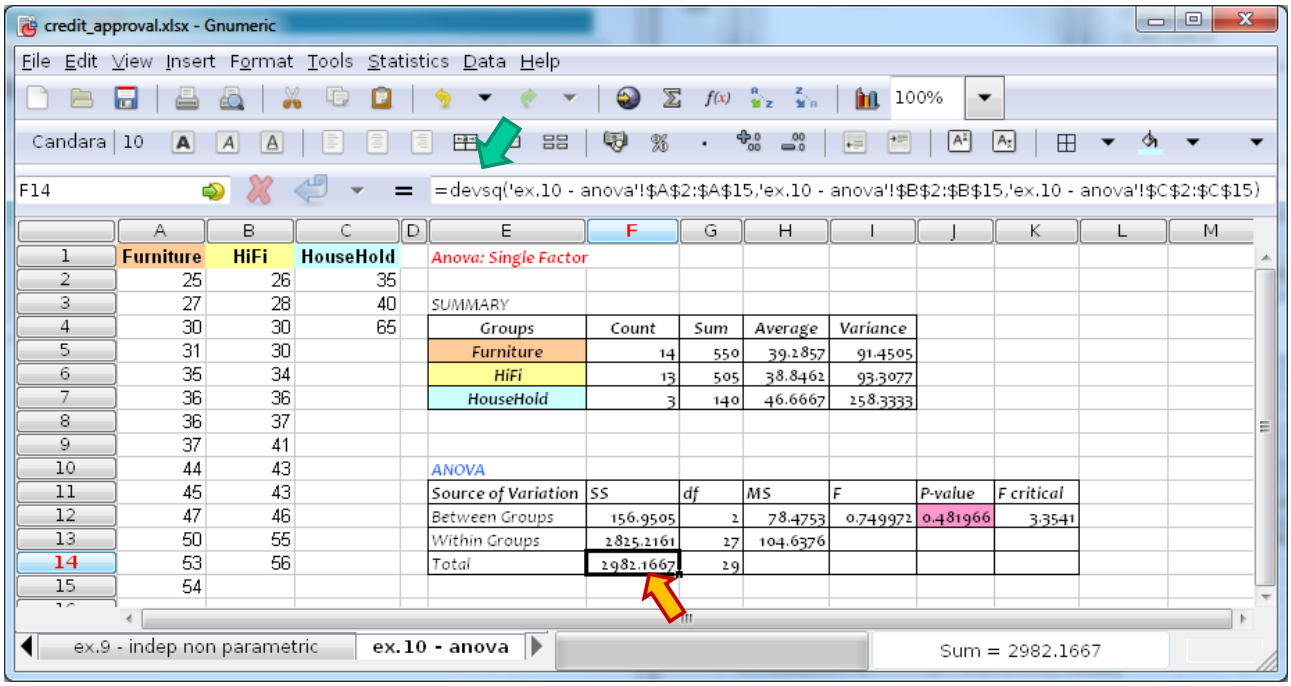


	A	B	C
1	Furniture	HiFi	HouseHold
2	25	26	35
3	27	28	40
4	30	30	65
5	31	30	
6	35	34	
7	36	36	
8	36	37	
9	37	41	
10	44	43	
11	45	43	
12	47	46	
13	50	55	
14	53	56	
15	54		
16			

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.



Gnumeric fournit les caractéristiques conditionnelles et le tableau d'analyse de variance.



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_Y	Attribute_X	Description				Statistical test		
		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
age	reason	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

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 (http://www.statsci.org/excel.html). Certaines, comme la librairie « matrix.xla », sont très performantes et passent haut la main les tests de précision (http://digilander.libero.it/foxes/). 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 (http://en.wikipedia.org/wiki/Gnumeric). 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, « Statistical Accuracy of Spreadsheet Software », 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 (Statistical Reference Datasets – 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 ».