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. « The Risks of Using Spreadsheets for Statistical Analysis », 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. SAS provides this kind of feature, Microsoft 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** (http://www.gnumeric.org/). 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

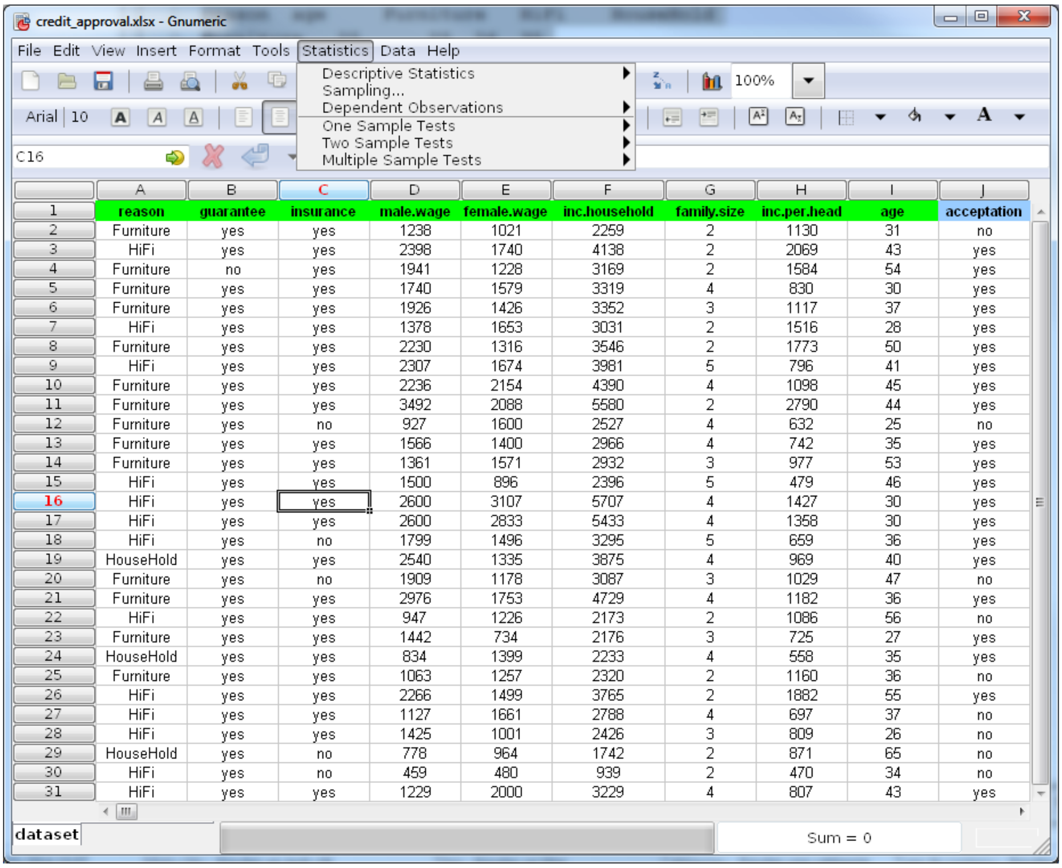
[1] The French version of this tutorial was written in **May 2014**. There was a version available for Windows on that date. That is no longer the case today (the last version with Windows binaries was 1.12.17, August 2014).

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

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



	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 – 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).

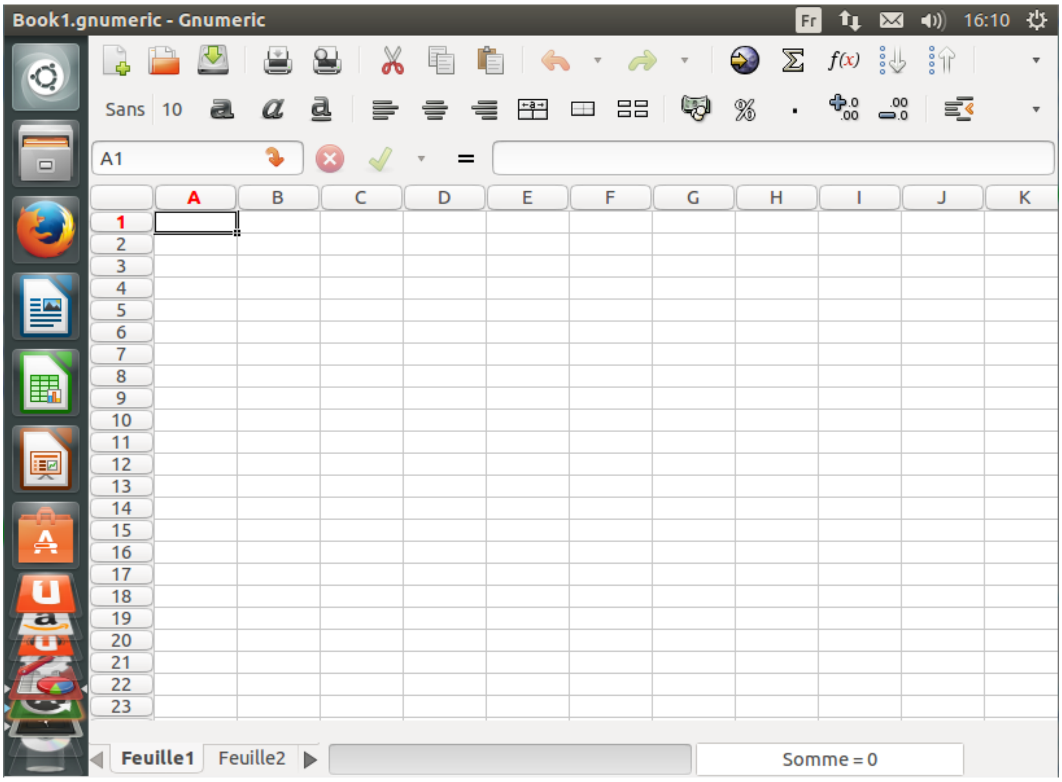


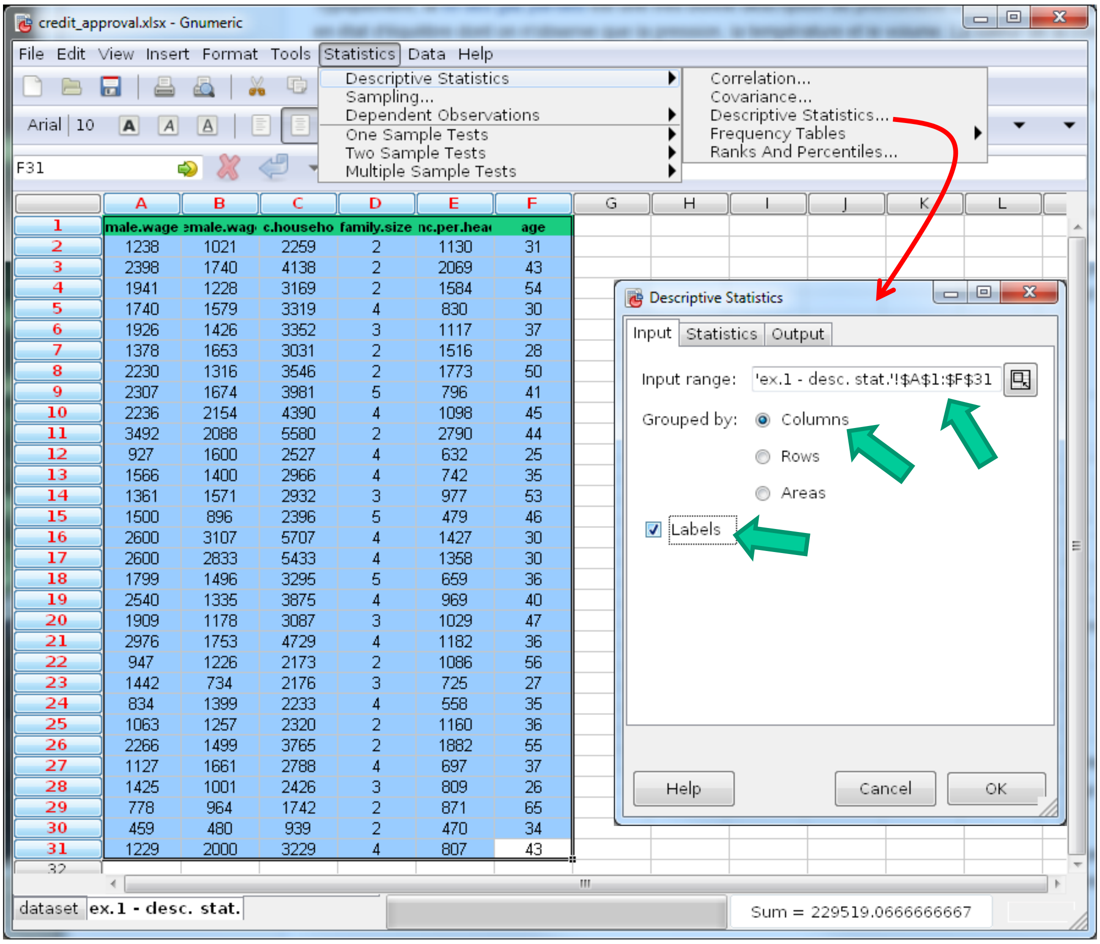
Figure 2 - Gnumeric under Ubuntu

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.

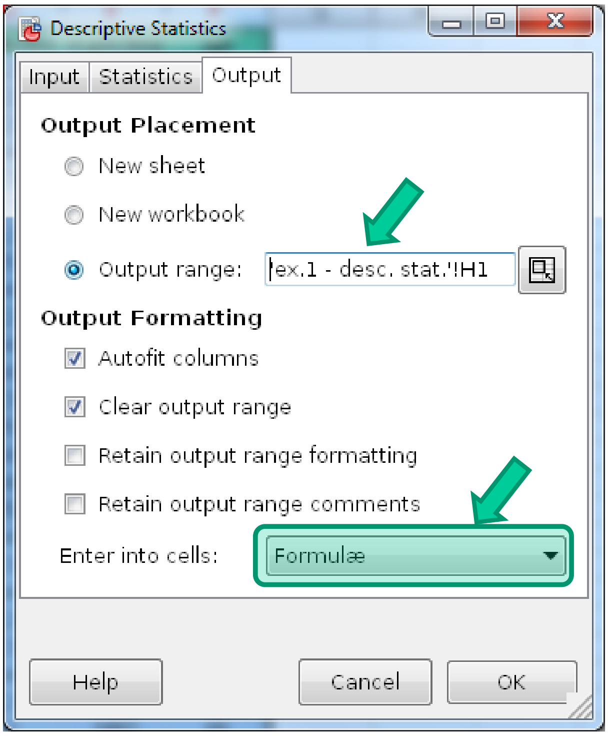
3 Statistical analysis with Gnumeric

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:

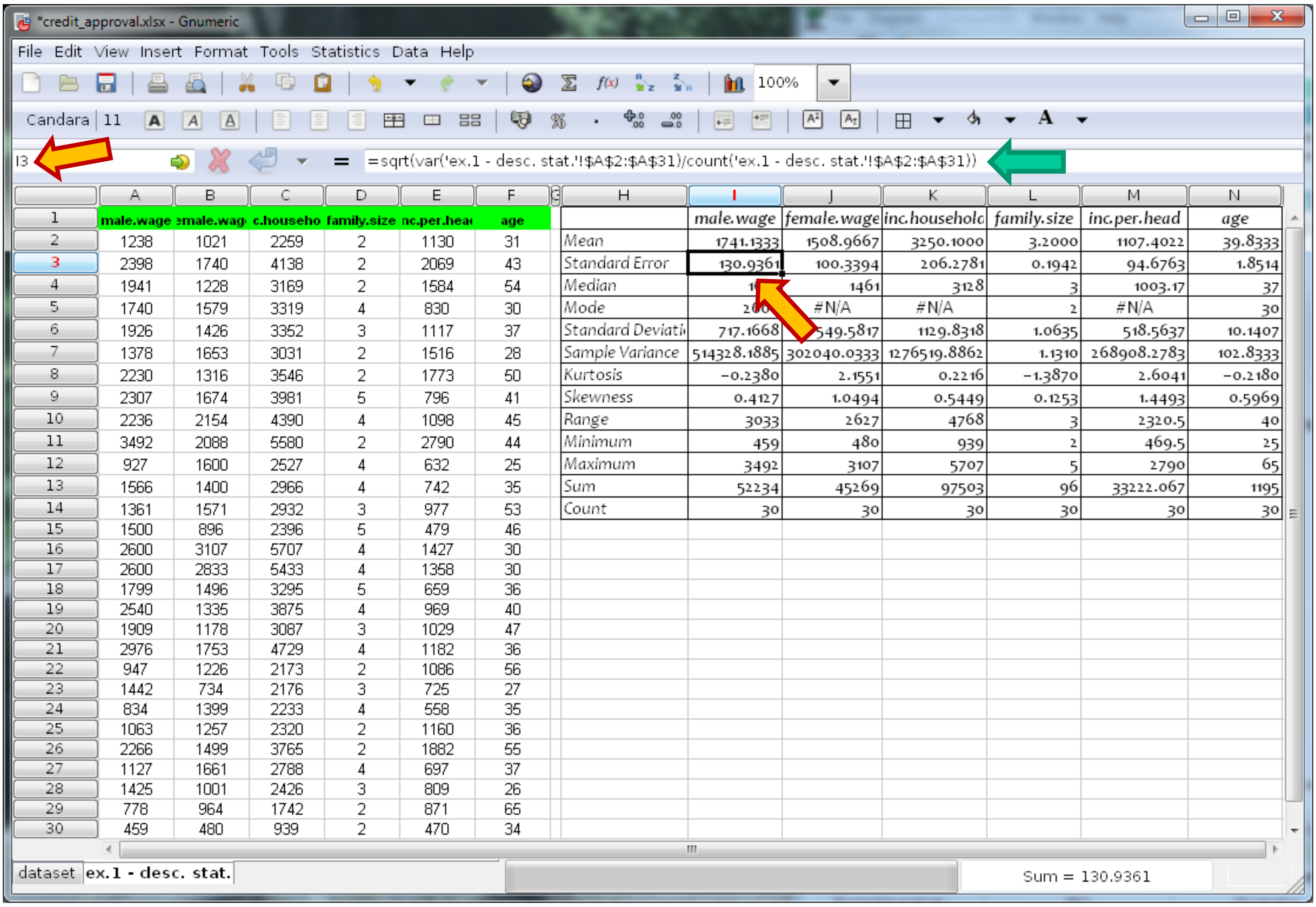


Into the INPUT tab, the data range must be selected. The variables are organized by columns in our case. We select the LABELS option to specify that the first row of the data range corresponds to the variable names.



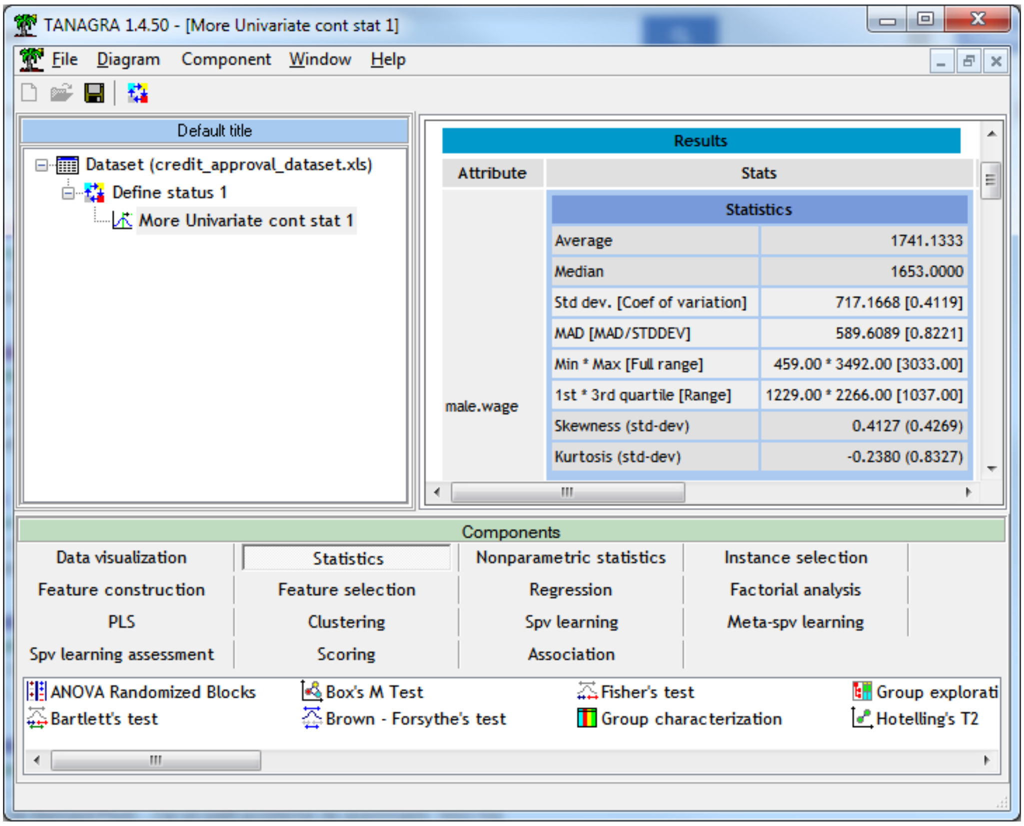
We do not change anything in the STATISTICS tab. In the OUTPUT tab, we specify the coordinates for the output. We observe the "Enter into cells: Formulae" option. It means that the results will be inserted as formulas. Thus, if the values into the data range are changed, the results will be automatically 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).



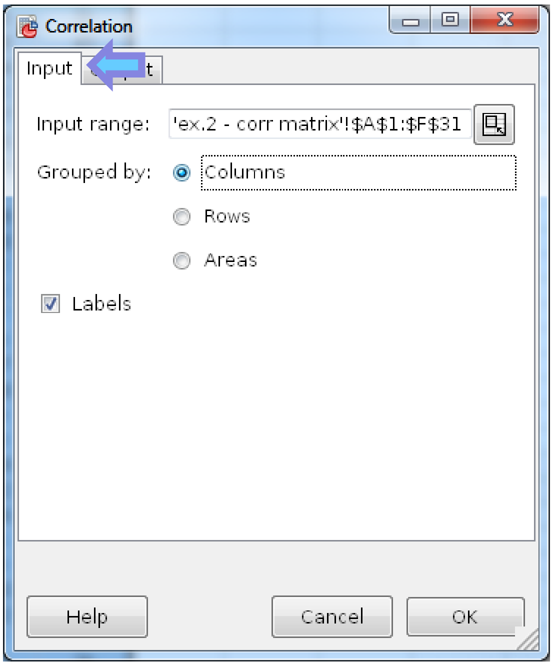
Let us see the standard deviation for the variable "X : male.wage". Into **I3** cell, we see the formulae $s_{\bar{x}} = \sqrt{\frac{s_x^2}{n}} = \sqrt{\frac{514328.1885}{30}} = 130.9361$. The estimated variance s_x^2 of X is into **I7**.

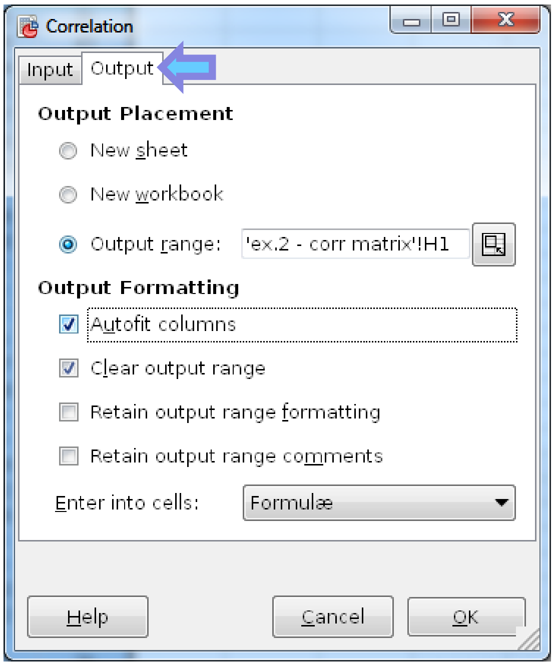
By comparison, we obtain the following results for "male wage" under **Tanagra 1.4.50**. The results are consistent.



3.2 Correlation tool

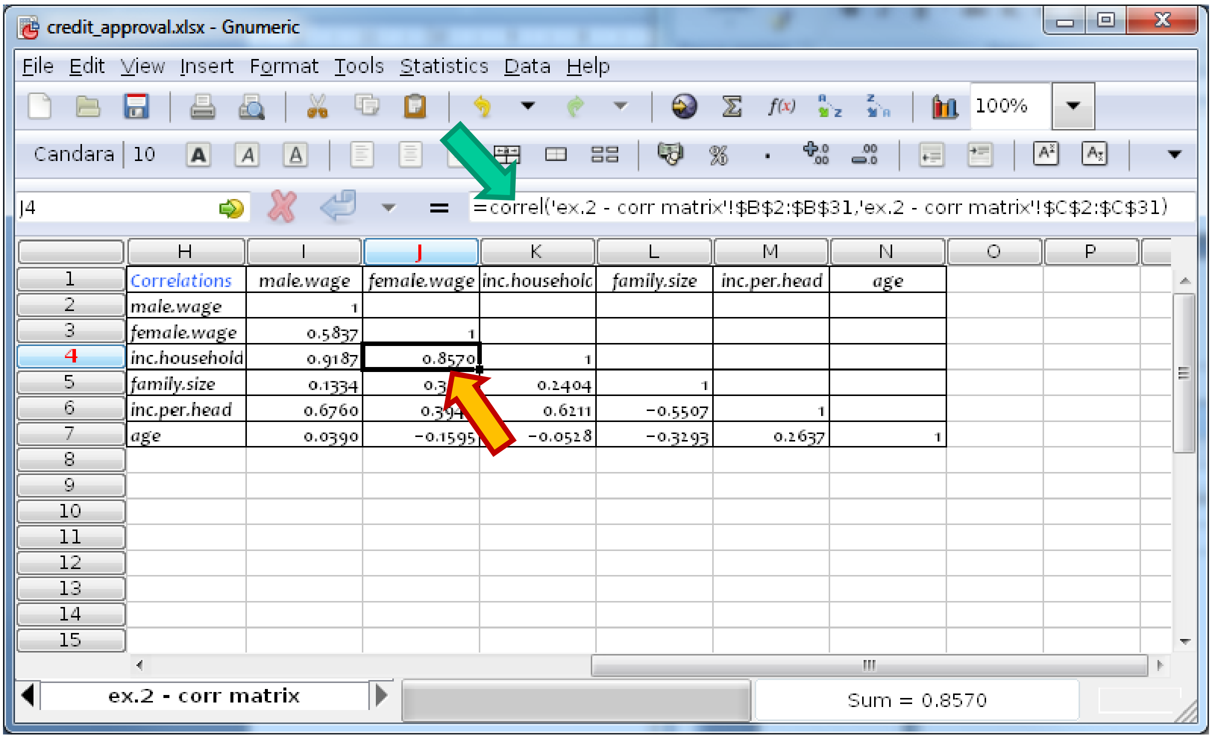
We use the same numeric variables to calculate the matrix of pairwise correlations coefficients. We duplicate the data range into a new worksheet **"ex.2 – corr matrix"**. After we select the data range, we click on the **Statistics / Descriptive Statistics / Correlation** menu.





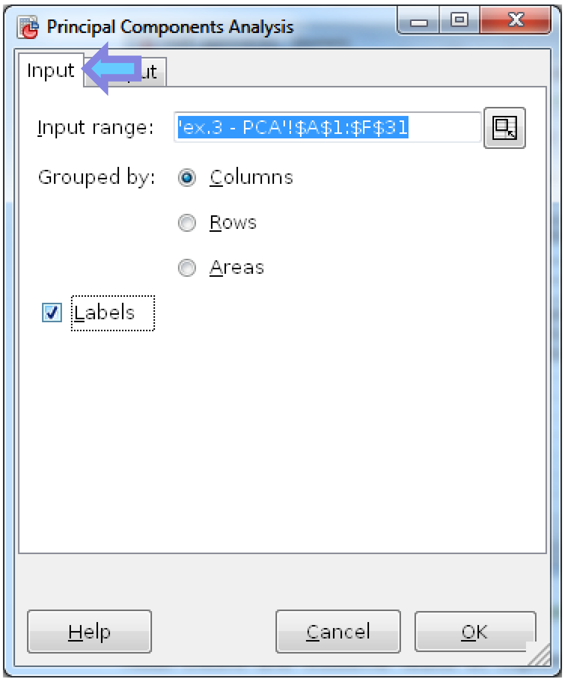
We check the input range and set the right parameters (variables grouped in columns, labels). Into the OUTPUT tab, we set the output range.

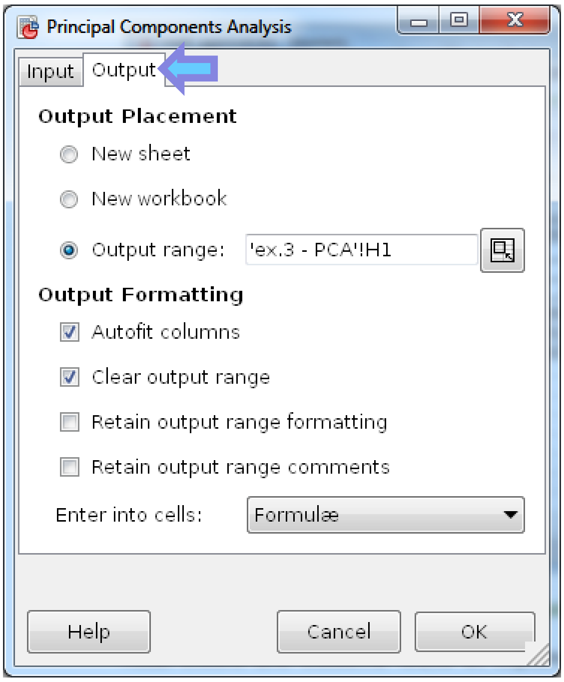
We note that the correlation coefficients are obtained with the **CORREL** function.



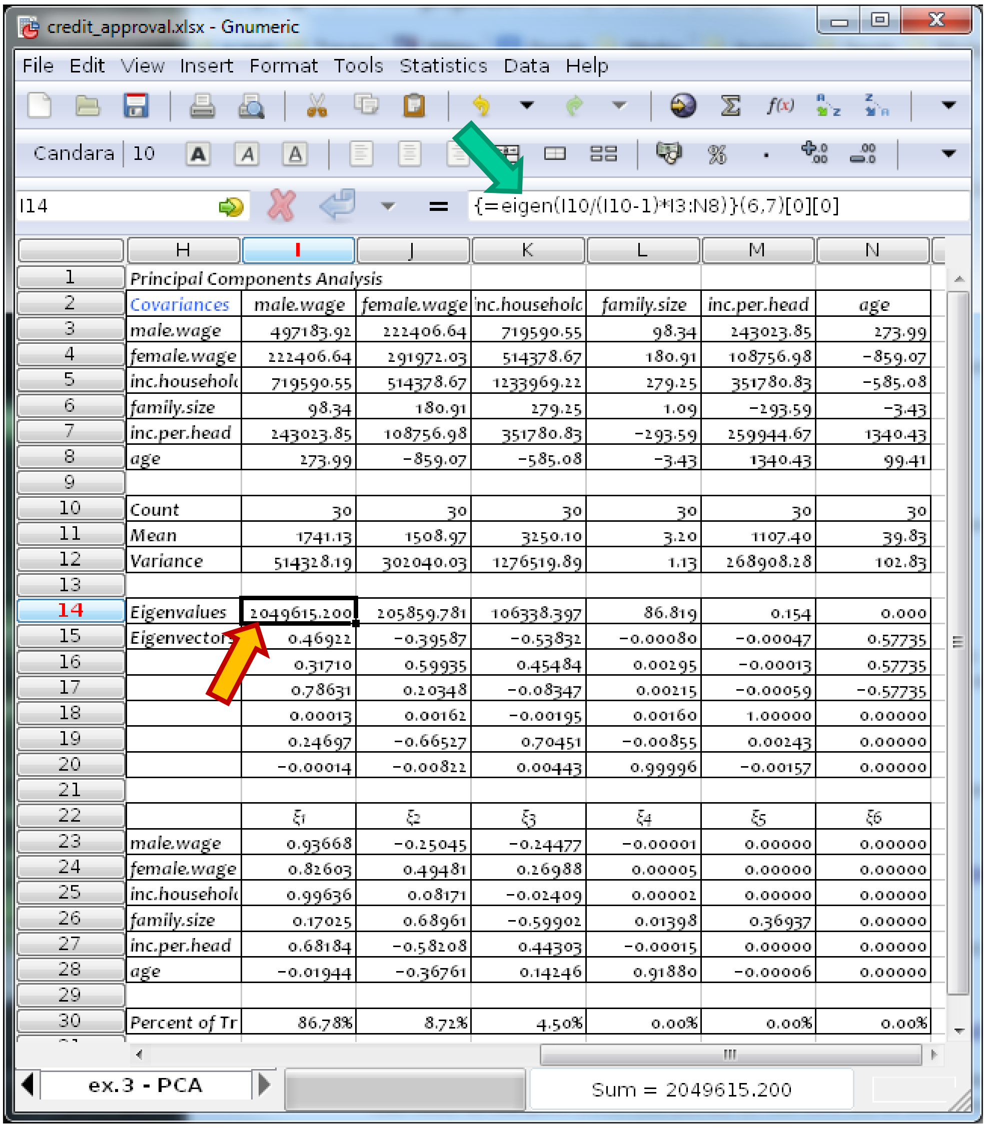
3.3 Principal component analysis (PCA)

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:





We obtain the following results:



credit_approval.xlsx - Gnumeric

I14 {=eigen(I10/(I10-1)*I3:N8)}(6,7)[0][0]

	H	I	J	K	L	M	N
1	Principal Components Analysis						
2	Covariances	male.wage	female.wage	inc.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.househol	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		ξ1	ξ2	ξ3	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 - 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):

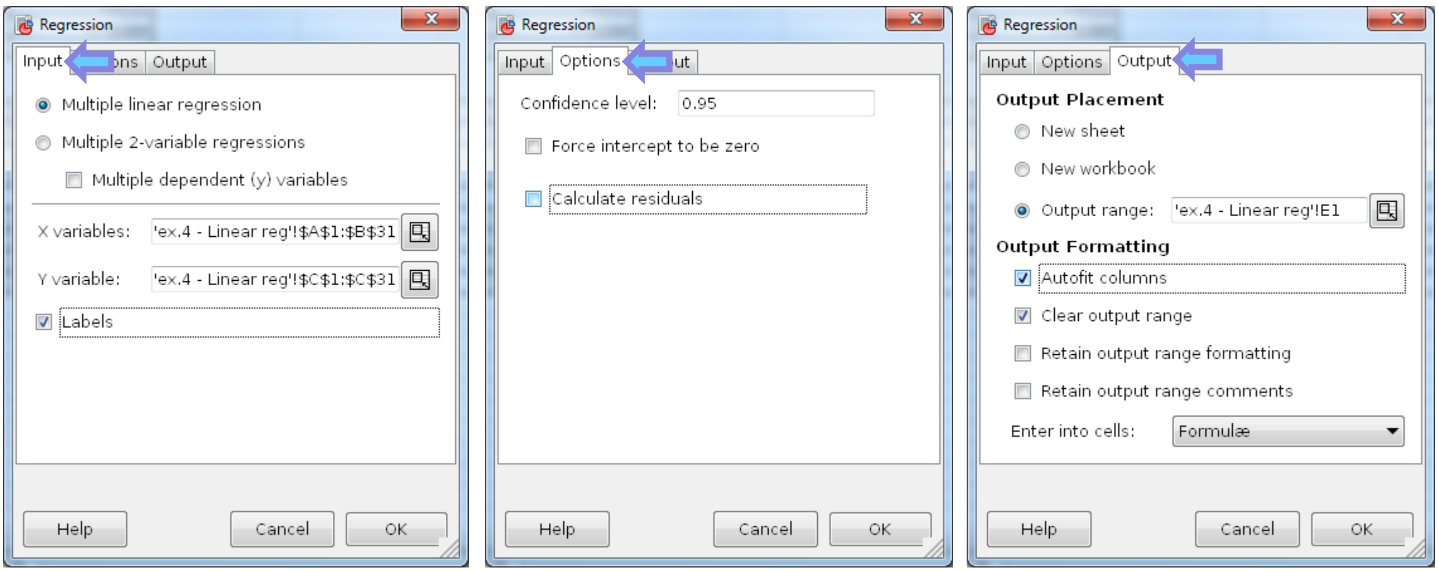
Axis	Eigen value	Difference	Proportion (%)	Cumulative (%)
1	1981294.694	1782296.905	86.78%	86.78%
2	198997.789	96204.006	8.72%	95.49%
3	102793.782	102709.858	4.50%	100.00%
4	83.925	83.776	0.00%	100.00%
5	0.149	0.149	0.00%	100.00%
6	0	-	0.00%	100.00%
Tot.	2283170.339	-	-	-

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 ($\frac{n}{n-1} \to 1$). 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.

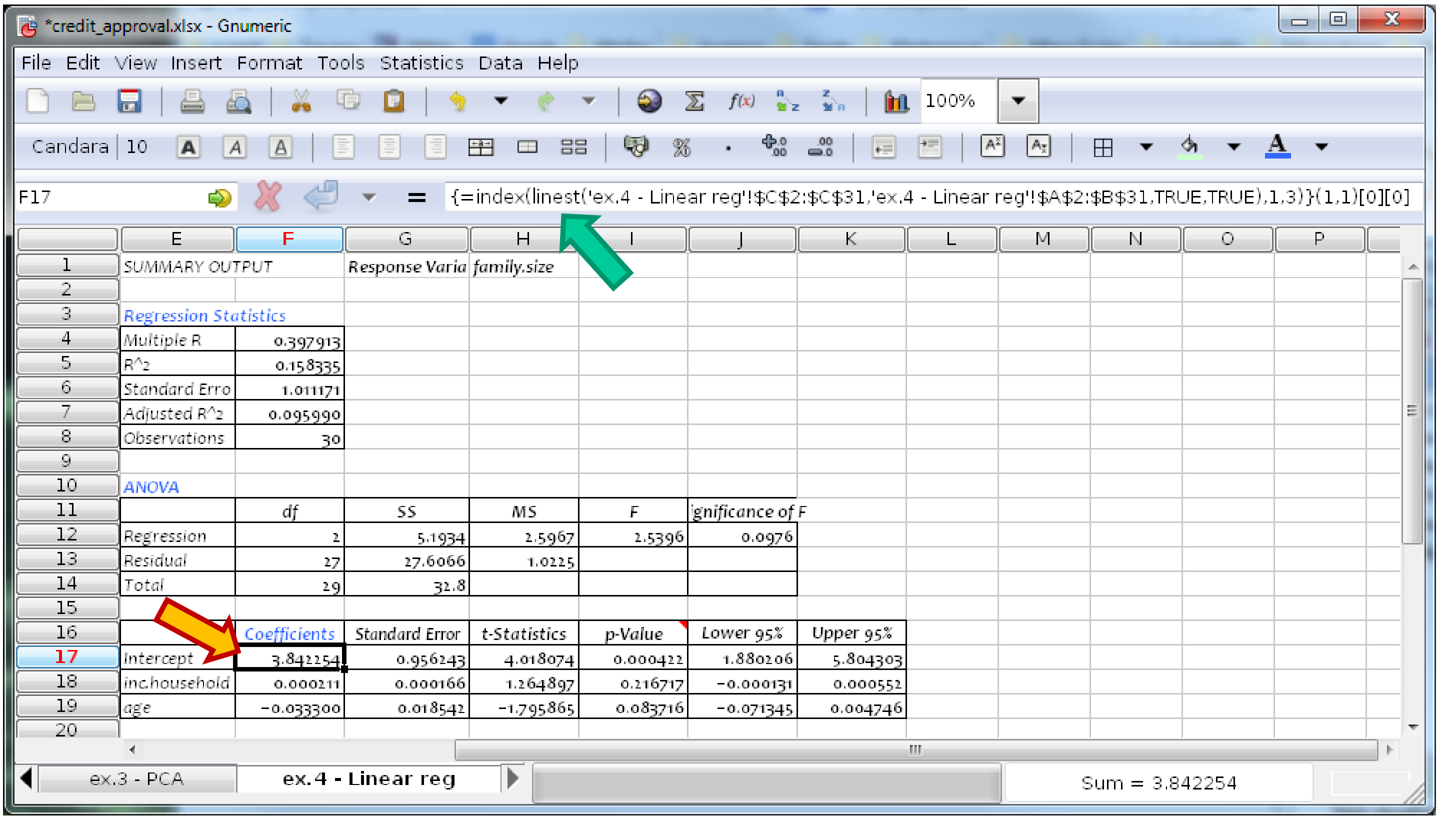
3.4 Linear regression

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:

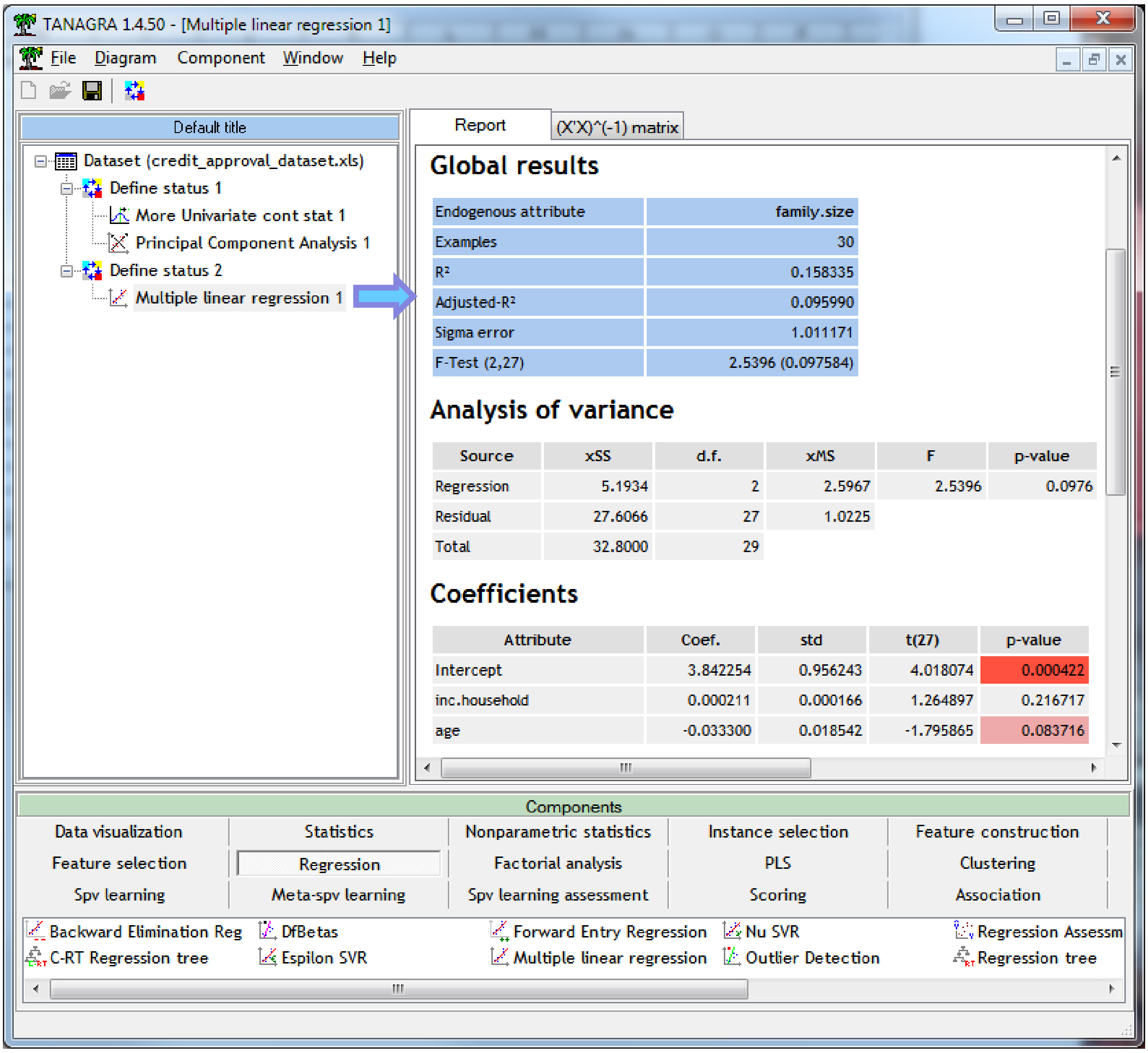


Y is the dependent variable (familiy.size), X is the range of the explanatory variables (inc.household and age). Gnumeric uses the LINEST function. It reorganizes the results for a presentation consistent with the standard statistical tools. It picks the different values in an internal table with the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



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.

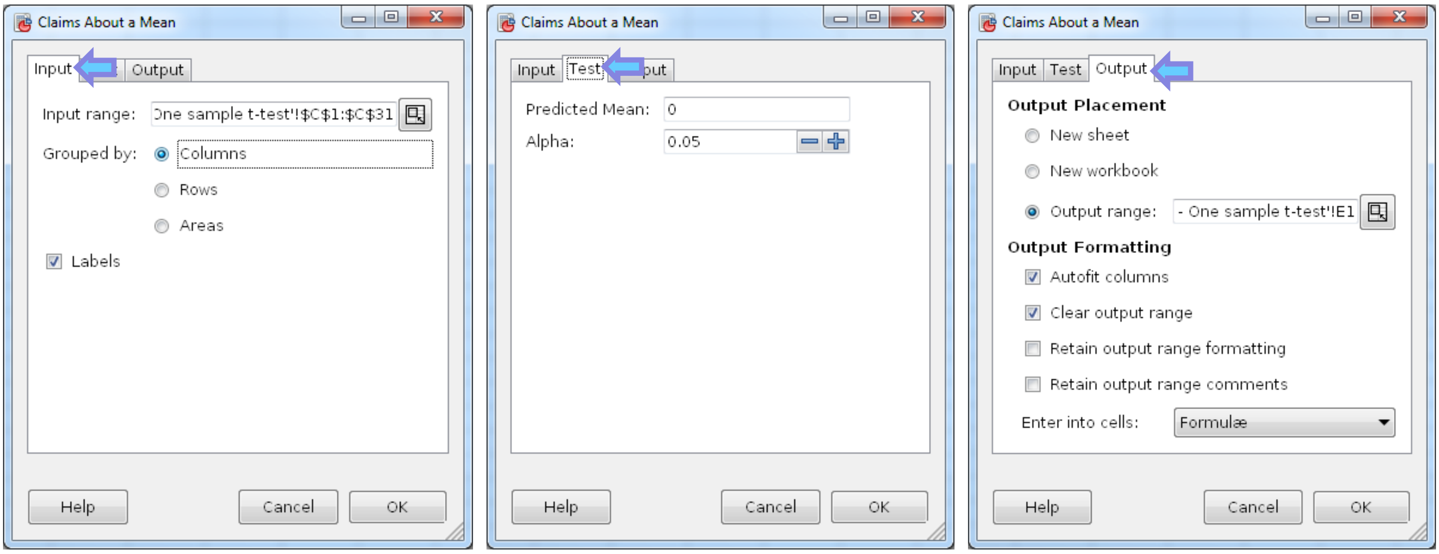
We obtain the same results with Tanagra. The organization is identical.



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.

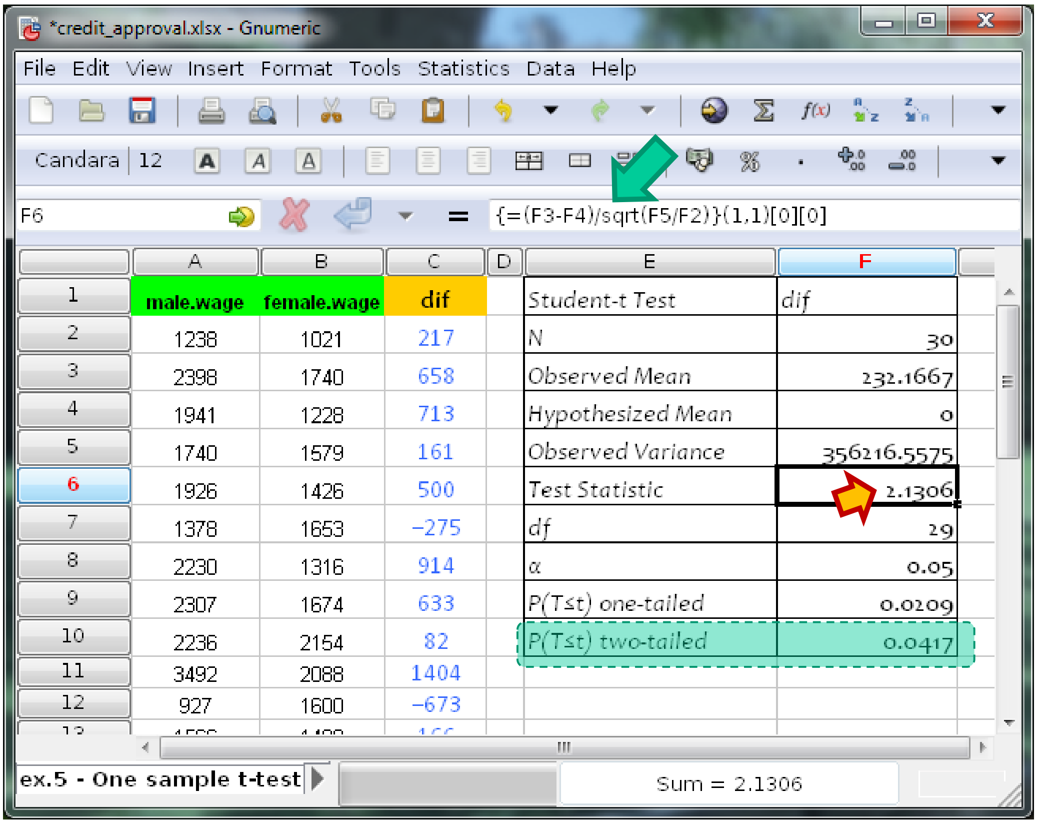
After we select the column DIF, we click on the **Statistics / One Sample Tests / Claims About a Mean** menu. We set the following parameters into the dialog setting.



At the 5%, we reject the null hypothesis[2]. We see at the cell **F6** the formulae ($\mu_0 = 0$ for our example, this the Predicted Mean option into the TEST tab):

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

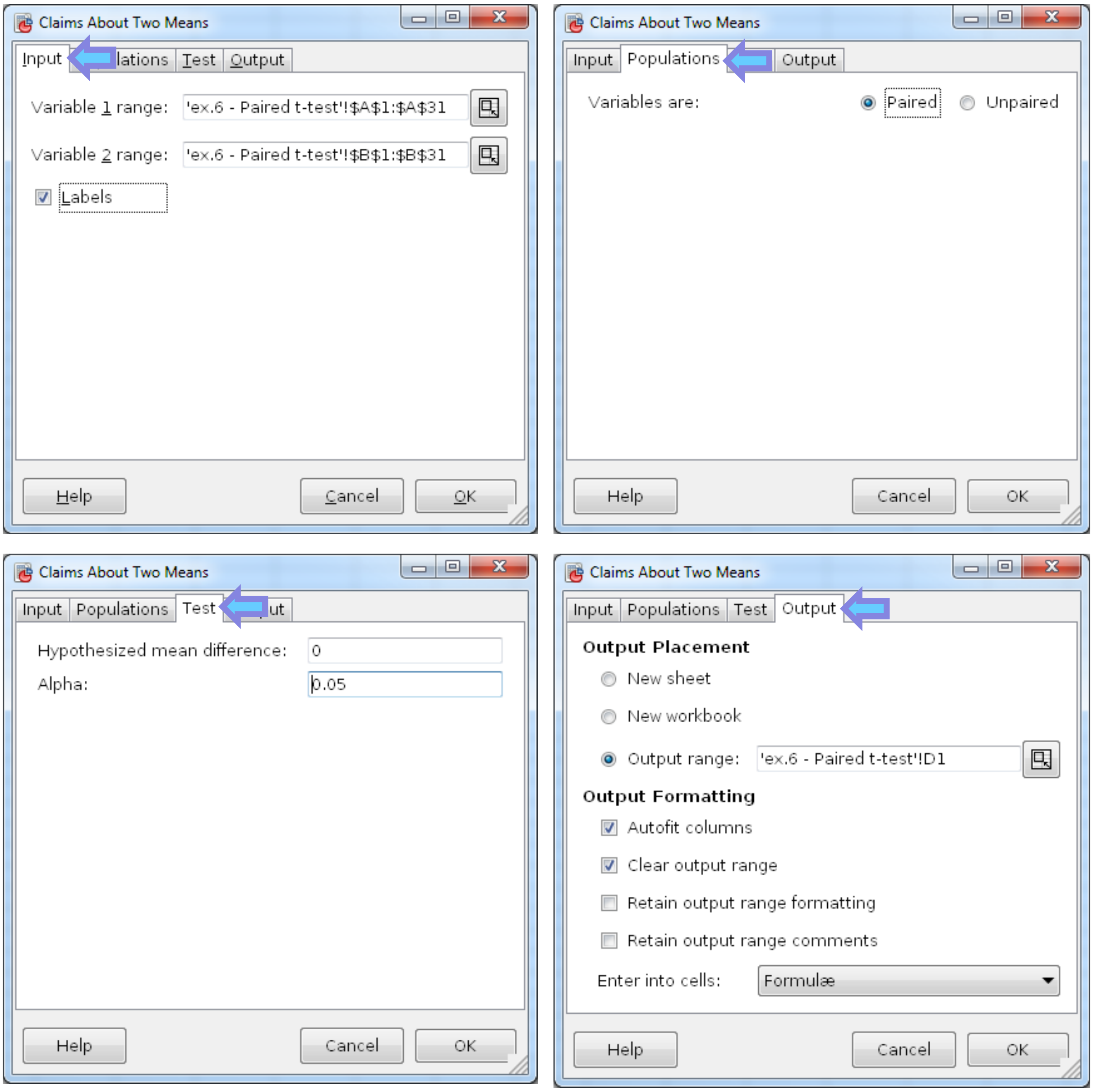
The p-value is obtained with the TDIST function.



[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**].

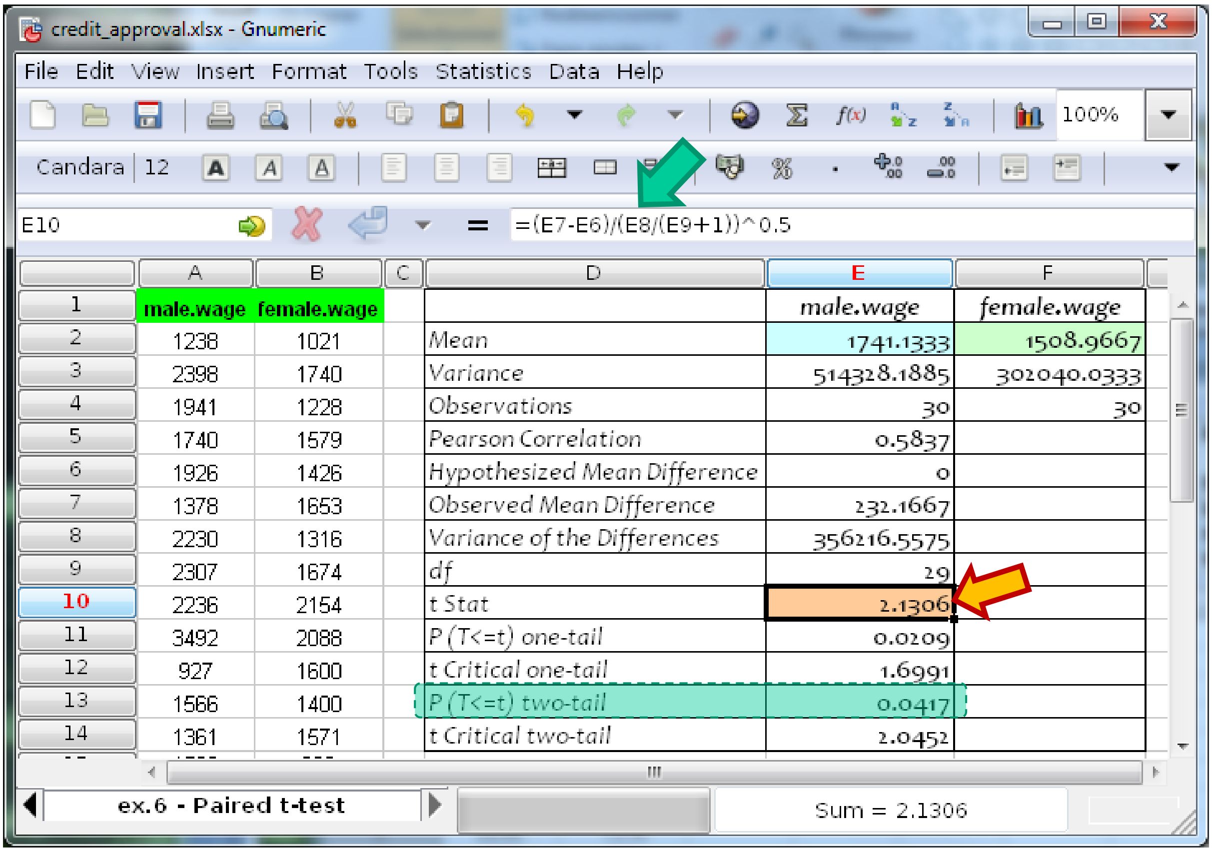
3.6 Comparison of two means – Paired samples

Another way to compare the male and female wages inside the household is to conduct a comparison of means between paired samples[3]. 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.

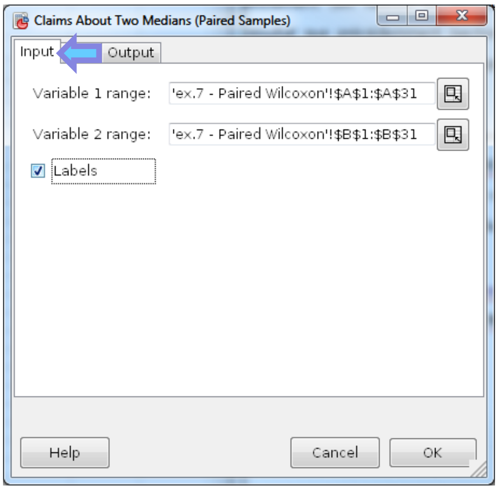
[3] https://onlinecourses.science.psu.edu/stat500/node/51

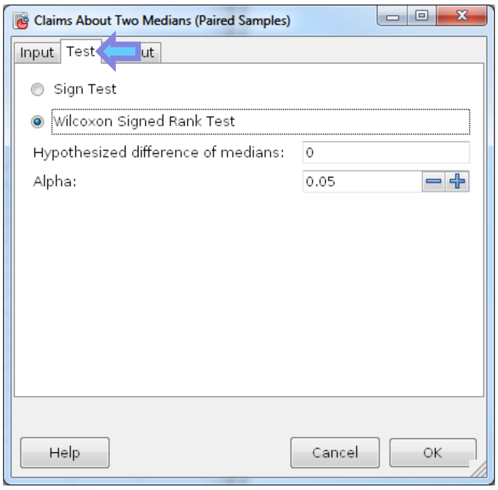


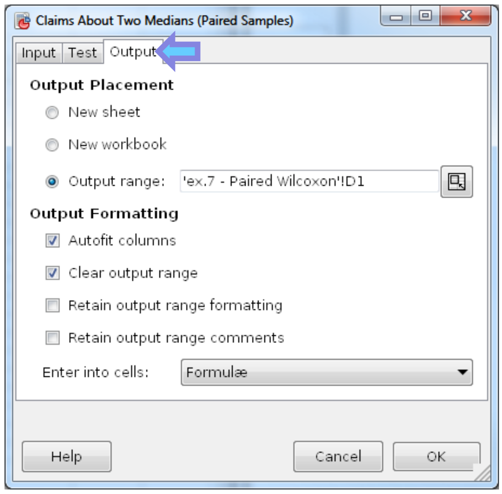
By a different process, we obtain exactly (the values and distributions of the test statistics are identical) the same result as previously (section **Erreur ! Source du renvoi introuvable.**). Men and women inside a household have not the same wages in average. It seems that men are advantaged.

3.7 Non-parametric test – Paired samples

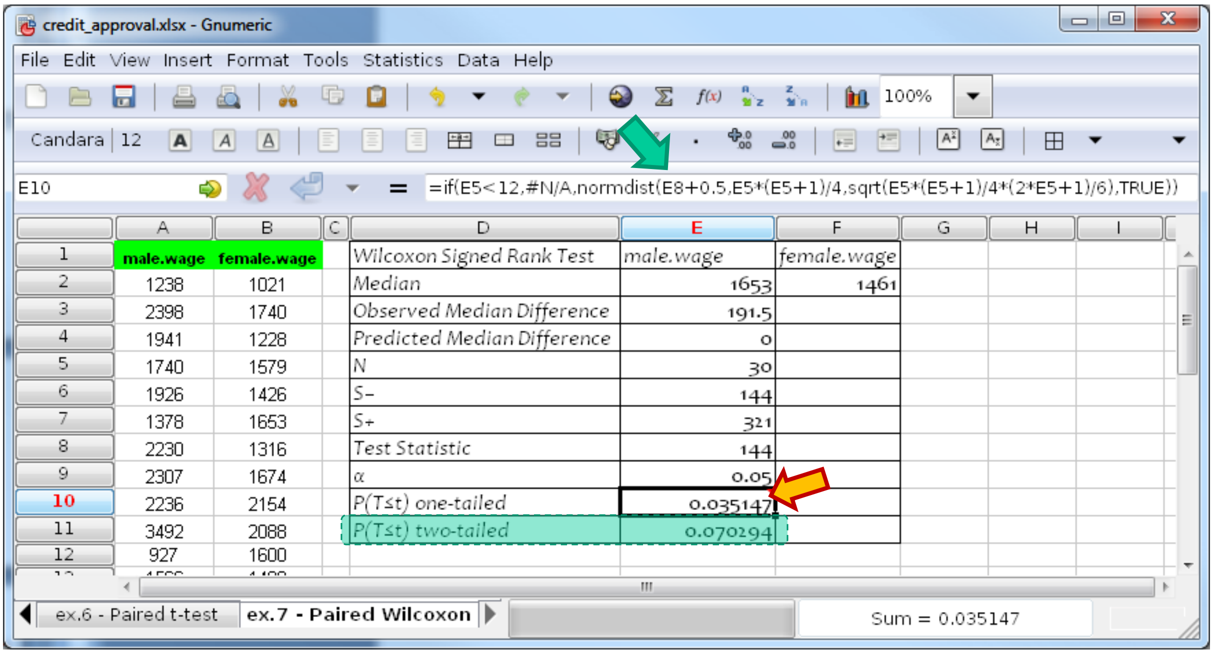
We can answer to this question - is the wages of the men and the women are equivalent inside the household - by the means of a non-parametric test. We use Wilcoxon signed rank test. The test statistic is based on the rank of the differences. We do not need that the populations are assumed to be normally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







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:



The sample size being enough (n ≥ 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		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

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

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

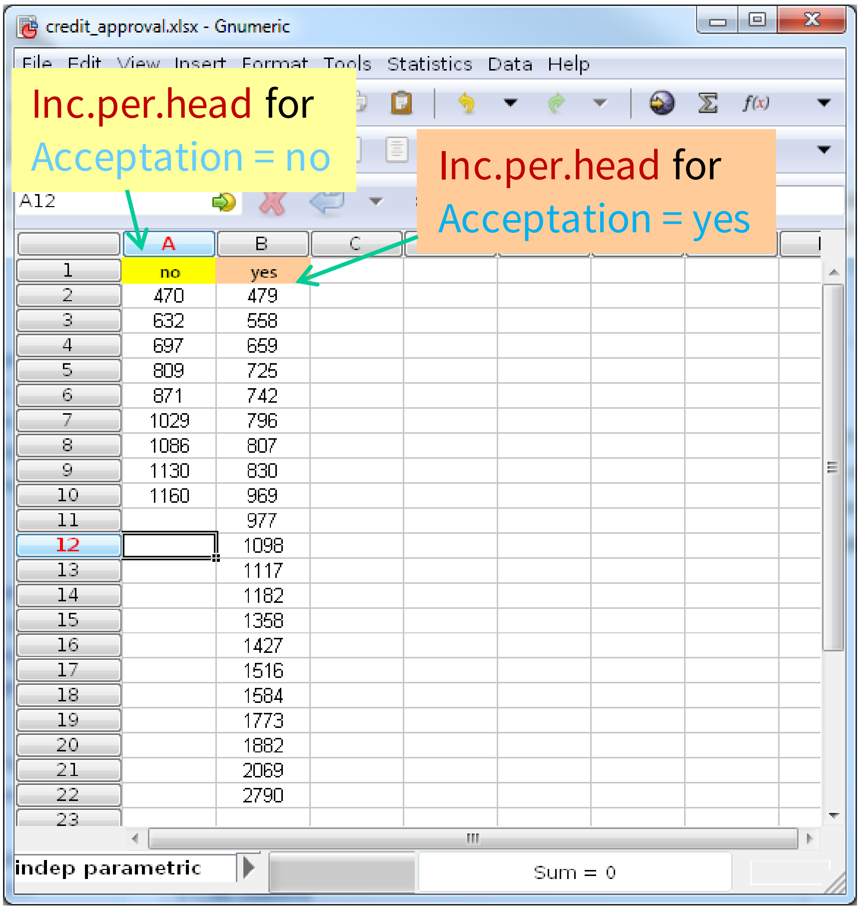
Using the cumulative distribution function $\Phi(.)$ of the standardized normal distribution,

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

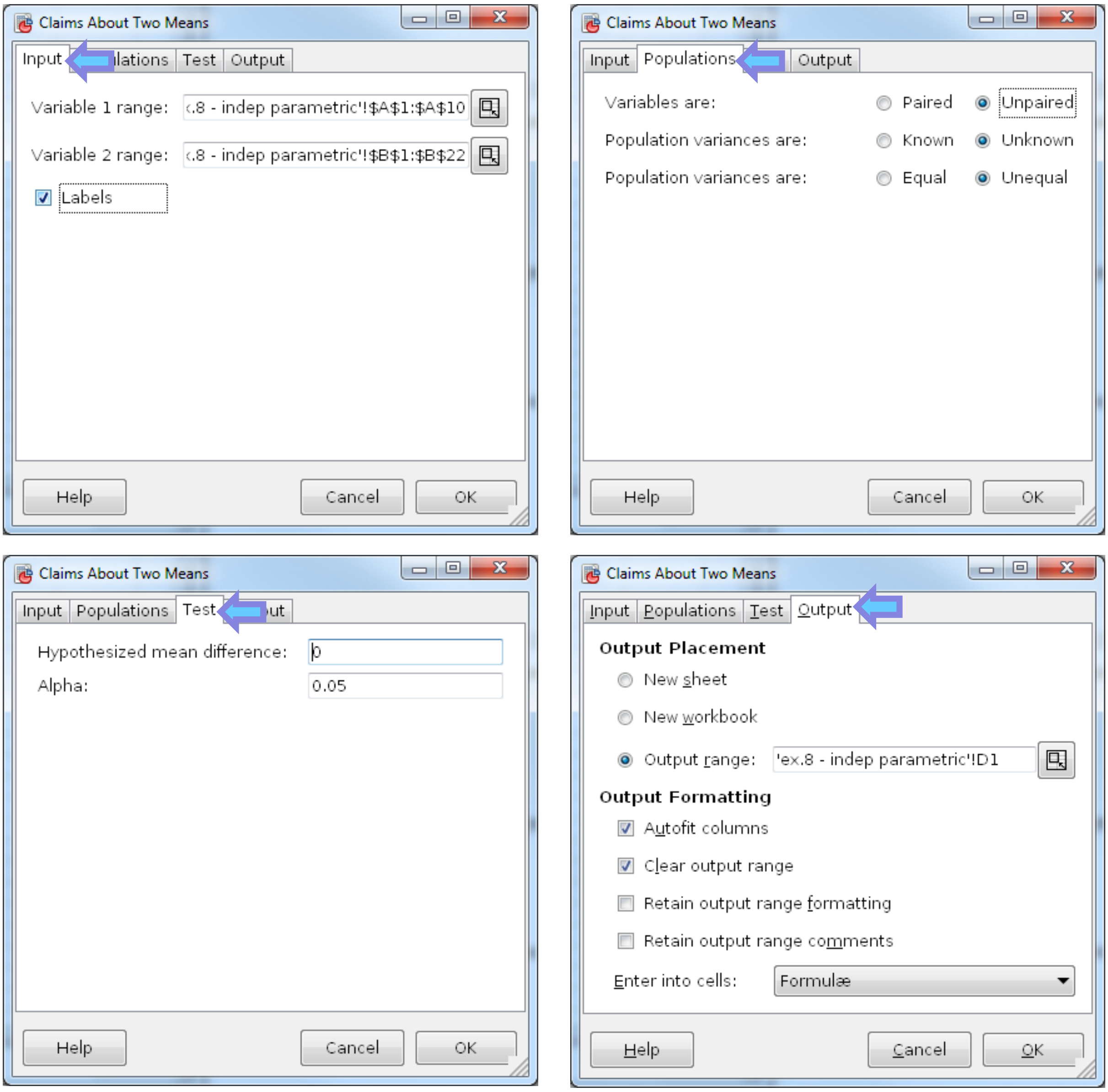
This is the p-value provided by Gnumeric.

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.

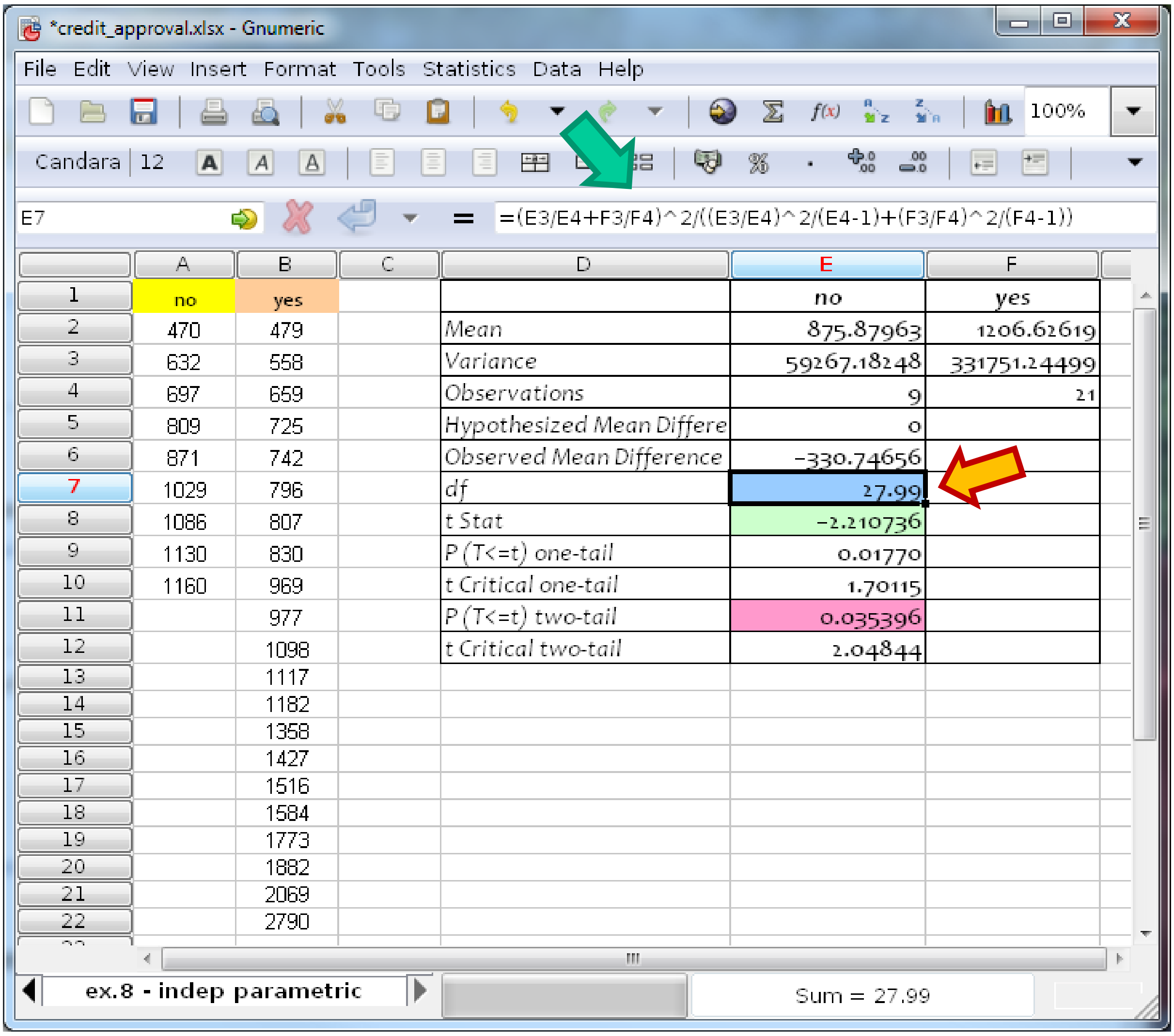


We see below the settings for the procedure **Statistics / Two Sample Tests / Claims about two means / Unpaired Samples, Unequal Variances**. We assume that the two populations have unequal variances.



The procedure is based on the Welch's t-test (http://en.wikipedia.org/wiki/Welch's_t_test). 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.



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

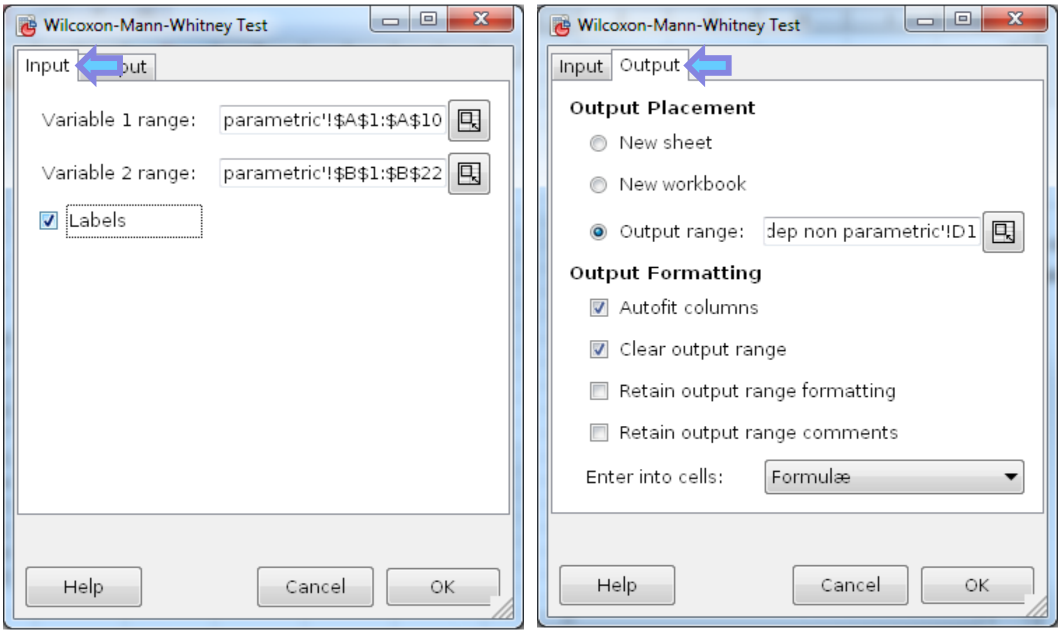
Attribute_Y	Attribute_X	Description				Statistical test	
		Value	Examples	Average	Std-dev	T	-330.7466 / 149.6092 = -2.210736
inc.per.head	acceptation	no	9	875.8796	243.4485	d.f.	27.99
		yes	21	1206.6262	575.9785	p-value	0.035393
		All	30	1107.4022	518.5637		

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.

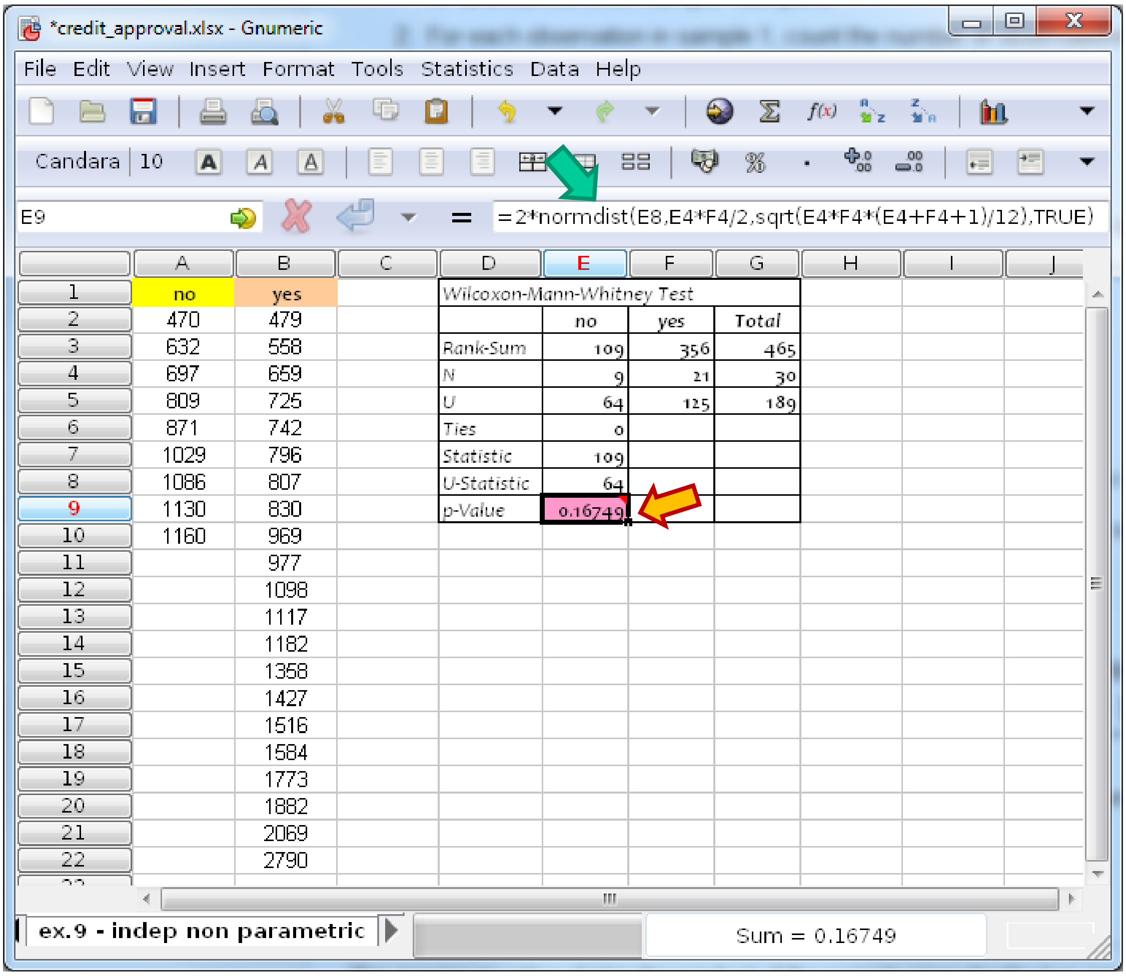
3.9 Non-parametric test – Independent samples

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.



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.



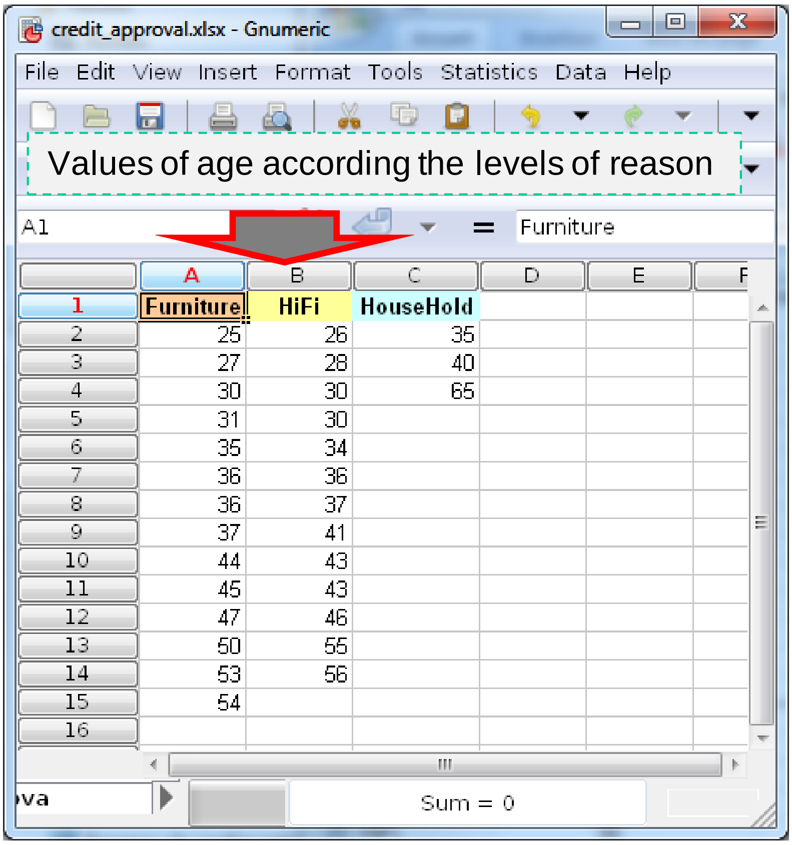
no	yes
470	479
632	558
697	659
809	725
871	742
1029	796
1086	807
1130	830
1160	969
	977
	1098
	1117
	1182
	1358
	1427
	1516
	1584
	1773
	1882
	2069
	2790

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

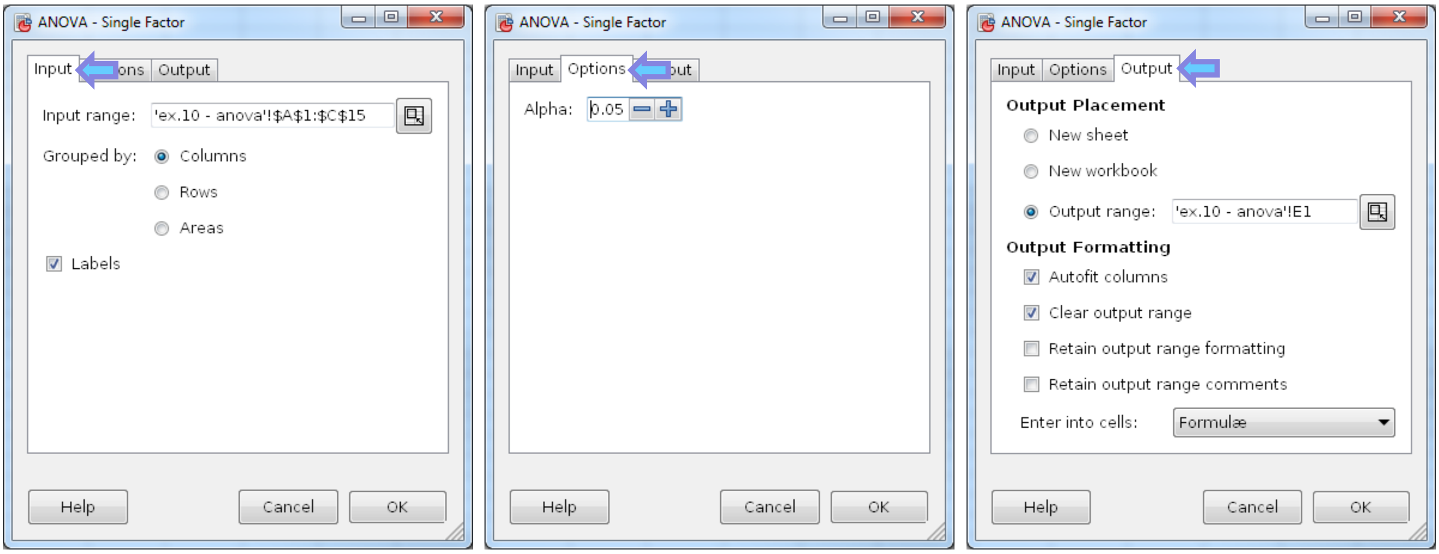
		Value	Examples	Average	Rank sum	Rank mean	Mann-Whitney U	64
inc.per.head	acceptation	no	9	875.8796	109	12.1111	E(U)	94.5
		yes	21	1206.6262	356	16.9524	V(U)	488.25
		All	30	1107.4022	465	15.5	|Z|	1.38032
							P(>|Z|)	0.16749

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

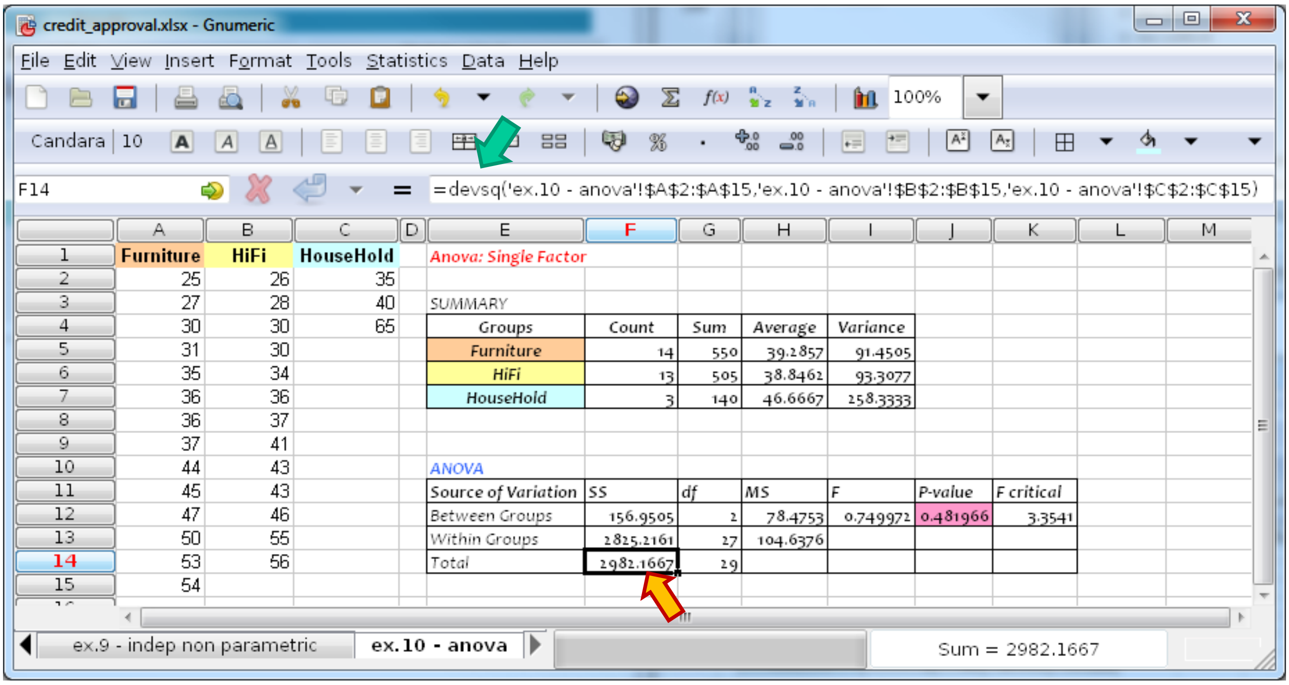
We want to compare the age of the persons according to the "reason" of the loan. We create a new sheet "**ex.10 – Anova**". We have a list of 3 columns because "reason" has 3 distinct possible values {furniture, hifi, household}. The number of observations can be different into the columns. Here is the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



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.



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



The function DEVSQ is essential in the calculations.

Results									
Attribute_Y	Attribute_X	Description				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, « Statistical Accuracy of Spreadsheet Software », 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 (Statistical 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 ».