1 Objectif

Présentation de l'add-in « Real Statistics » pour Excel (http://www.real-statistics.com/).

Excel – je dirais plutôt le tableur de manière générique - est un des outils favoris des « data scientist ». Les sondages Kdnuggets sur la question le confirment[1]. Il arrive systématiquement dans les 3 premiers logiciels les plus utilisés ces dernières années. Les raisons de ce succès ont été maintes fois évoquées sur ce blog : il est très répandu, tout le monde sait le manipuler, c'est un instrument puissant pour la mise en forme et la préparation des données.

Excel est moins à son avantage lorsqu'il s'agit d'effectuer des calculs statistiques. D'aucuns pointent du doigt son manque de précision et la relative pauvreté de sa bibliothèque de fonctions statistiques et d'analyse de données[2,3]. Les add-ins (ou add-on, macros complémentaires) semblent alors constituer une solution privilégiée pour associer les calculs spécialisés aux fonctionnalités usuelles des tableurs.

Les add-ins agissent de différentes manières. Certains établissent un pont en transférant simplement les données vers les logiciels de data mining qui opèrent alors indépendamment, en stand-alone. C'est le cas de la macro « tanagra.xla » pour Tanagra[4], ou du dispositif RExcel pour R[5]. D'autres logiciels fonctionnent en sous-main, de manière transparente, après réception des données et renvoient les résultats dans une feuille de calcul ou dans une fenêtre dédiée. C'est le cas de des add-ins de SAS[6] et de SQL Server[7]. D'autres enfin intègrent directement les traitements, et procèdent aux calculs en les programmant en VBA (Visual Basic pour Applications) ou en les incorporant dans des DLL (librairies compilées) externes.

La librairie « **Real Statistics** » du Dr Charles Zaiontz appartient à cette troisième catégorie. C'est une solution simple comme je les aime. La copie d'un fichier « RealStats-2007.xlam » (pour la version 2007 d'Excel) suffit pour disposer pleinement de toutes les fonctionnalités. Il

[1] KDnuggets Polls, « Analytics / Data Mining software used ? », May 2013 ; « KDnuggets 15th Annual Software Poll : RapidMIner continues to lead », June 2014.

[2] K. Keeling, R. Pavur, « Statistical Accuracy of Spreadsheet Software », The Amercial Statistician, 65:4, 265-273, 2011.

[3] IBM SPSS Statistics, « The risks of Using Spreadsheets for Statistical Analysis ». On sera un peut plus circonspect concernant cet article. Rédigé et publié par un éditeur de logiciel de statistique, on pourrait croire qu'il n'est pas dénué d'arrière-pensées ; les références utilisées sont anciennes, on imagine qu'Excel a évolué positivement depuis.

[4] « L'add-in Tanagra pour Excel 2007 et 2010 », Août 2010. La liaison est unidirectionnelle mais, les sorties de Tanagra étant en HTML, il est possible de les copier dans une feuille de calcul Excel.

[5] « Connexion entre R et Excel via RExcel », Décembre 2011. Notons que la connexion joue dans les deux sens, il est possible, via le même dispositif, de récupérer des objets R dans Excel.

[6] « SAS add-in 4.3 pour Excel », Avril 2012.

[7] Microsoft, Data Mining Add-ins. Voir un exemple d'utilisation sur un site de partage de vidéos célèbre.

n'y a pas d'installation fastidieuse à réaliser, avec des bibliothèques à tiroirs que l'on est obligé de chercher à droite et à gauche. La macro complémentaire se suffit à elle-même, elle ne repose pas sur une DLL compilée. Grâce à cette autonomie, il a été possible de multiplier les versions pour les différentes configurations d'Excel (des add-ins existent pour Excel 2013, 2010, versions antérieures à Excel 2003, version pour Mac)[8]. Les résultats des calculs statistiques sont insérés dans les feuilles de calculs sous forme de formules s'appuyant sur des fonctions standards d'Excel (ex. les opérations matricielles, nous pouvons ainsi retracer les étapes des traitements) ou de nouvelles fonctions spécifiques intégrées dans la librairie, que nous pouvons appeler directement dans d'autres feuilles de calculs. Il y a donc deux manières d'utiliser l'add-in : soit, comme nous le ferons dans ce tutoriel, exploiter les boîtes de dialogue dédiées permettant de spécifier les données à traiter et paramétrer les méthodes ; soit en appelant directement les nouvelles fonctions disponibles.

« Real Statistics » est une excellente librairie, à conseiller aux personnes qui souhaitent travailler exclusivement dans l'environnement Excel pour réaliser les traitements statistiques. Elle est d'autant plus intéressante qu'elle est accompagnée d'une documentation particulièrement riche, permettant de comprendre dans le détail la teneur de chaque méthode. Nous décrivons dans ce tutoriel le mode opératoire de l'add-in et, dans certains cas, nous comparons les résultats avec ceux de **Tanagra 1.4.50**.

2 Chargement et installation de la librairie

L'add-in « Real Statistcs » est accessible sur le web (Figure 1). Plusieurs pages retiennent notre attention :

- La page de téléchargement. Plusieurs variantes relatives aux versions d'Excel sont disponibles (http://www.real-statistics.com/free-download/real-statistics-resource-pack/).
- Une page décrivant l'installation de la ressource dans Excel (http://www.real-statistics.com/free-download/real-statistics-resource-pack/#install). Il faut absolument la lire attentivement parce que l'utilisation initiale n'est pas évidente pour les non-initiés.
- Une page contenant des classeurs exemples (http://www.real-statistics.com/free-download/real-statistics-examples-workbook/).
- Les pages détaillant les traitements sous-jacents aux méthodes. Leur contenu pédagogique est particulièrement intéressant (ex. la description du test Box's M - http://www.real-statistics.com/multivariate-statistics/boxs-test-equality-covariance-matrices/boxs-test-basic-concepts/, avec notamment les transformations suivant la loi du

[8] http://www.real-statistics.com/free-download/real-statistics-resource-pack/

KHI-2 utilisée par Tanagra, et la loi de Fisher fournie par « Real Statistics » en rapport avec les caractéristiques des données ; voir section 4.7).

- La liste des fonctions portées par la librairie (http://www.real-statistics.com/excel-capabilities/supplemental-functions/), utilisables dans tout classeur Excel.

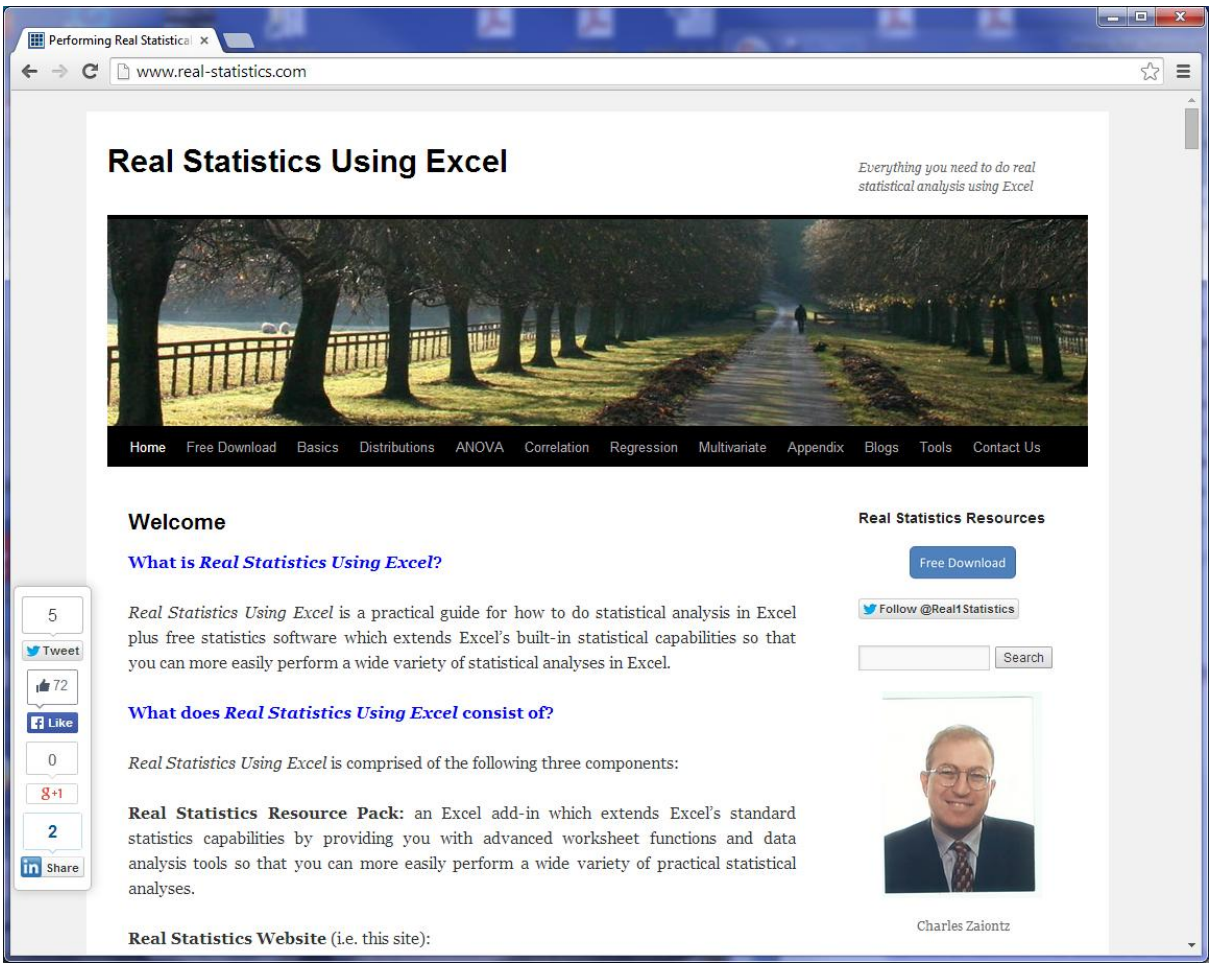


Figure 1 - Site web de la librairie "Real Statistics"

3 Données

Nous traitons le fichier utilisé lors de la présentation du tableur Gnumeric dans ce document. Il décrit n = 30 demandeurs de crédits à l'aide de p = 9 variables, 5 quantitatives et 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).

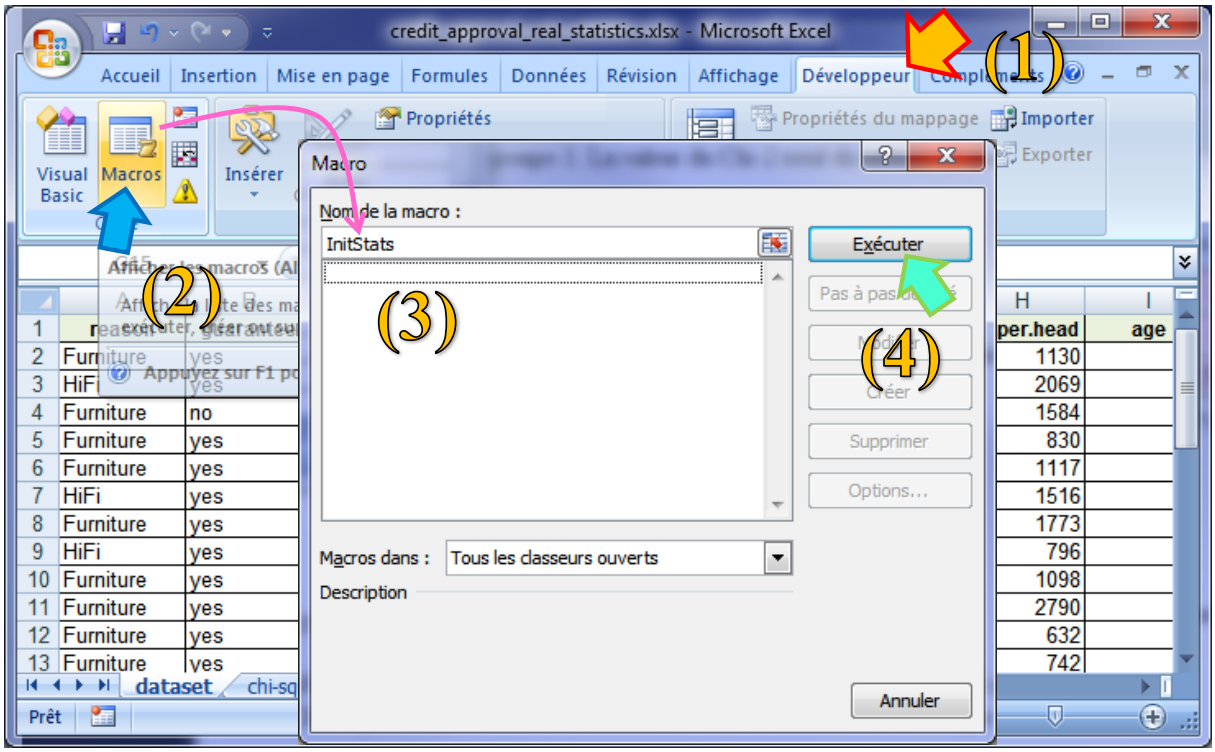
Voici les 5 premières lignes du classeur « **credit_approval_real_statistics.xlsx** ».

reason	guarantee	insurance	male.wage	female.wage	inc.household	family.size	inc.per.head	age	acceptation
Furniture	yes	yes	1238	1021	2259	2	1130	31	no
HiFi	yes	yes	2398	1740	4138	2	2069	43	yes
Furniture	no	yes	1941	1228	3169	2	1584	54	yes
Furniture	yes	yes	1740	1579	3319	4	830	30	yes
Furniture	yes	yes	1926	1426	3352	3	1117	37	yes

4 Traitements avec Real Statistics

Dans ce qui suit, nous décrivons méthodes statistiques proposées par la librairie « Real Statistics ». Nous détaillons les sorties en mettant parfois en contrepoint celles de Tanagra. A chaque analyse correspond une feuille de calcul distincte dans le classeur Excel.

Pour afficher la boîte de dialogue de démarrage, nous activons l'onglet Développeur (**1**), nous actionnons le bouton « MACROS » (**2**). Nous introduisons la commande InitStats dans la fenêtre de lancement (**3**) et nous cliquons sur le bouton « Exécuter » (**4**)[9]. **Remarque** : Avec la version **2.15**, un menu est maintenant disponible dans l'onglet « Compléments » ; nous pouvons également lance la boîte de démarrage avec le raccourci CTRL+M.



Une boîte de sélection des traitements statistiques apparaît. Nous pouvons choisir dans la liste le type d'analyse que nous souhaitons mener.

[9] Sur son site web, l'auteur décrit comment créer, une fois pour toutes, un raccourci dans le ruban de menus d'Excel afin d'éviter cette manipulation qui peut se révéler fastidieuse à la longue. Voir http://www.real-statistics.com/excel-capabilities/supplemental-data-analysis-tools/accessing-supplemental-data-analysis-tools/ ; section « **Quick Access Toolbar** ».

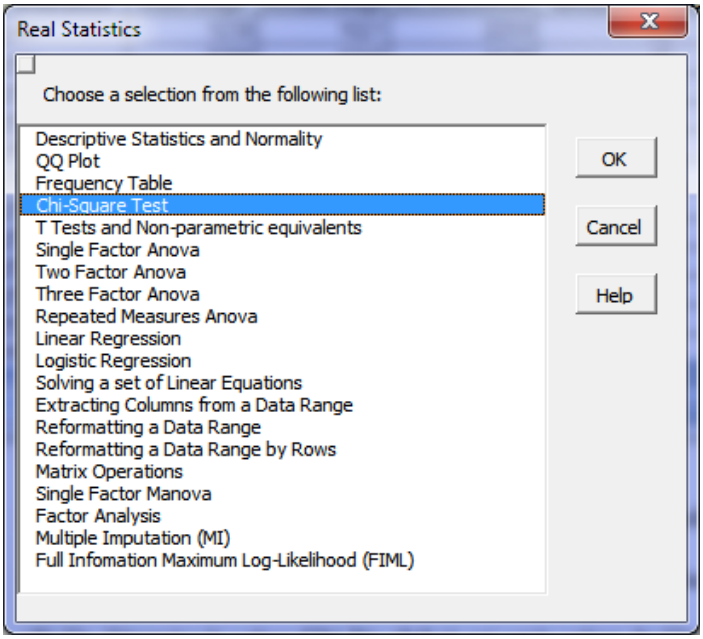
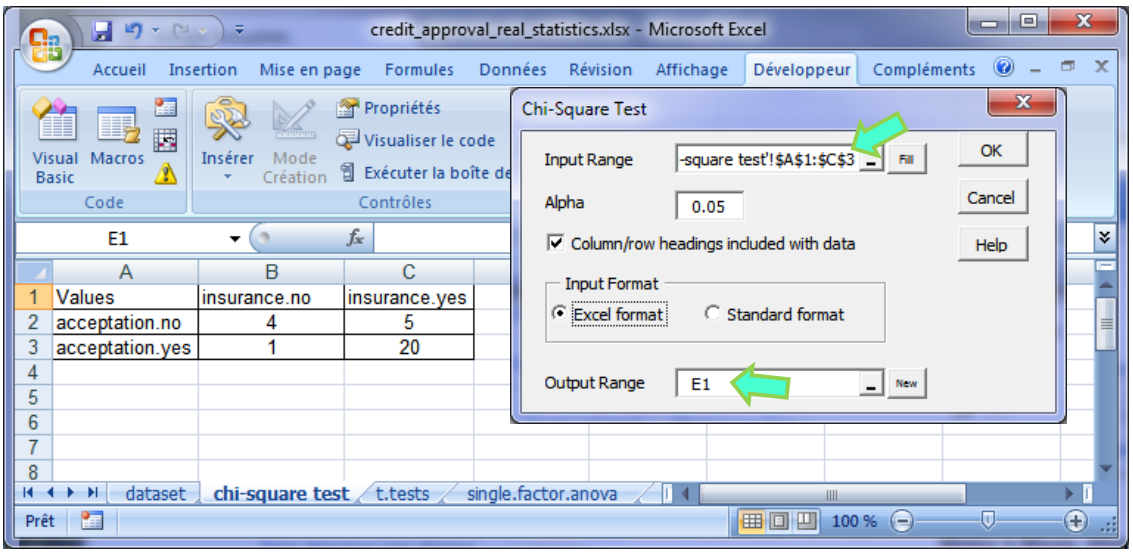


Figure 2 - Liste des méthodes activables interactivement avec Real Statistics

4.1 Test d'indépendance du Khi-2

Nous mettons en œuvre un test d'indépendance du Khi-2 en croisant les variables ACCEPTATION et INSURANCE (feuille « **chi-square test** »). Nous devons dans un premier temps former le tableau de contingence. Puis, nous sélectionnons, dans la fenêtre de lancement des méthodes, le « **Chi-Square Test** ». Une boîte de dialogue apparaît, nous spécifions la plage de données (**A1…C3**), et la coordonnée de la plage de sortie (on se contente d'indiquer le coin en haut et gauche de la plage, **E1**).



4.1.1 Lecture des résultats

Real Statistics nous fournit : le tableau sous l'hypothèse d'indépendance (**Expected Values**) ; un rapide diagnostic du tableau de contingence (effectifs, nombres de lignes et de colonnes)

(**SUMMARY**) ; et deux versions de la statistique du Khi-2, celle de Pearson et celle du rapport de vraisemblance. La normalisation de Cramer est affichée (**CHI-SQUARE**).

Expected Values

Values	insurance.no	insurance.yes	Total
acceptation.no	1.5	7.5	9
acceptation.yes	3.5	17.5	21
Total	5	25	30

Chi-Square Test

SUMMARY		Alpha	0.05
Count	*Rows*	*Cols*	*df*
30	2	2	1

CHI-SQUARE

	chi-sq	*p-value*	*x-crit*	*sig*	*Cramer V*	*Odds Ratio*
Pearson's	7.1429	0.0075	3.8415	yes	0.4880	16
Max likelihood	6.6277	0.0100	3.8415	yes	0.4700	16

A titre de comparaison, voici les valeurs proposées par le composant CONTINGENCY CHI-SQUARE de TANAGRA. Les valeurs présentées sont absolument cohérentes.

Row (Y)	Column (X)	Statistical indicator	
		Stat	Value
acceptation	insurance	d.f.	1
		Tschuprow's t	0.487950
		Cramer's v	0.487950
		Phi²	0.238095
		Chi² (p-value)	7.14 (0.0075)
		Lambda	0.333333
		Tau (p-value)	0.2381 (0.0086)
		U(R/C) (p-value)	0.1808 (0.0100)

Cross-tab

	yes	no	Sum
no	5	4	9
yes	20	1	21
Sum	25	5	30

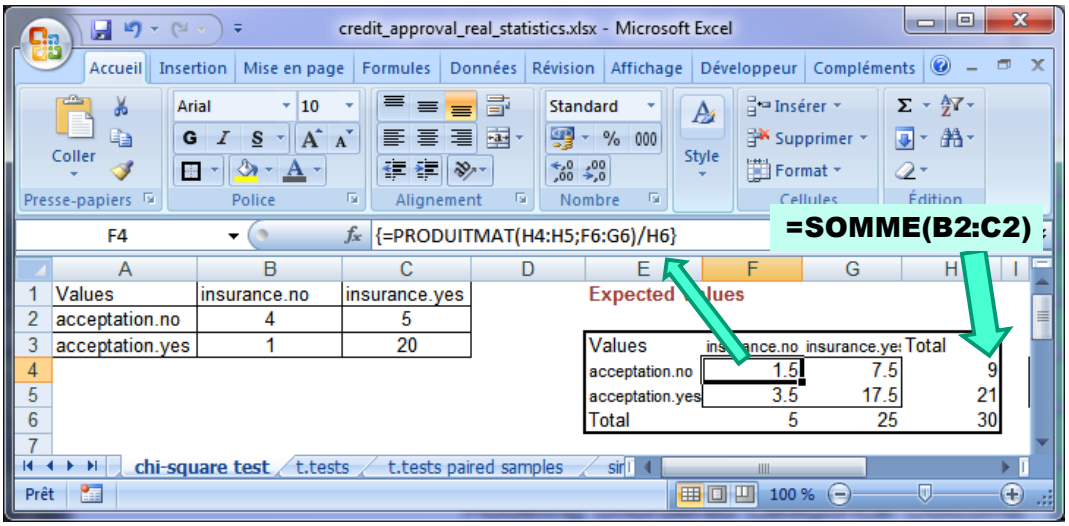
La statistique du maximum de vraisemblance est accessible avec le composant THEIL U.

Theil's U (Uncertainty Coefficient) for nominal attributes

Y	X	Theil's U	Chi²	d.f.	p-value	sigma	95% C.I.
acceptation	insurance	0.180829	6.63	1	0.0100	0.134790	-0.0834 ; 0.4450

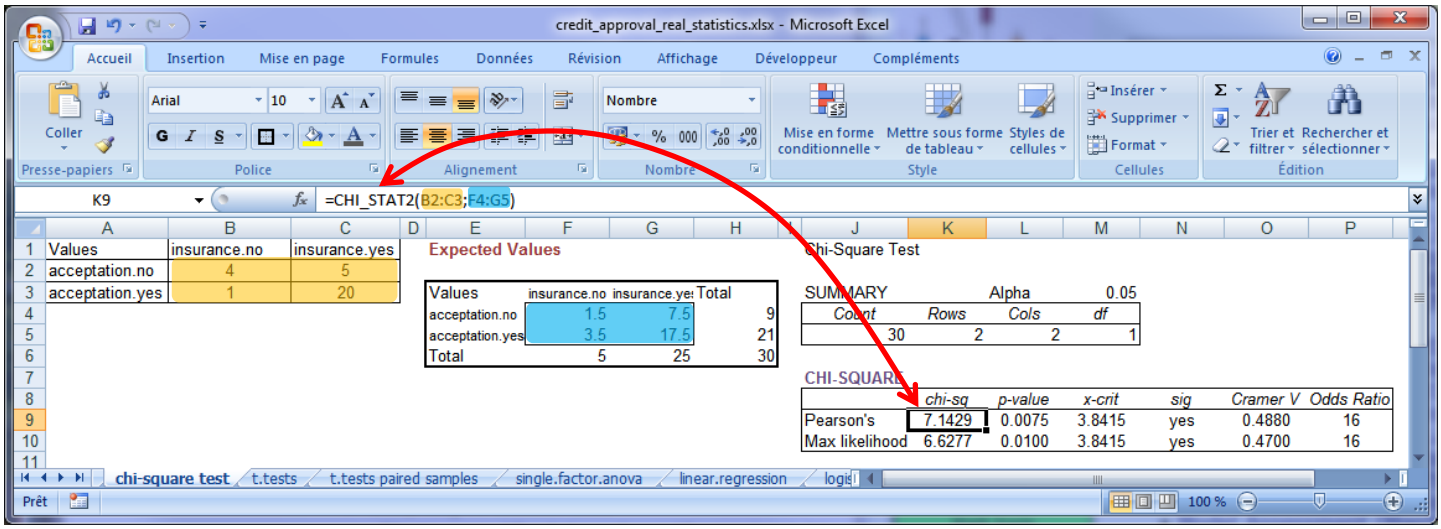
4.1.2 Détail des calculs avec Real Statistics

L'énorme intérêt de Real Statistics réside dans la possibilité d'accéder aux formules directement insérées dans la feuille de calcul Excel. Reprenons notre exemple ci-dessus.



En **H4** a été insérée la formule ''=SOMME(B2 :C2)'' permettant d'obtenir les effectifs marginaux ; en **F4**, le calcul matriciel produit les effectifs sous indépendance. Real Statistics s'est chargé de les insérer, mais nous aurions pu le faire nous-mêmes manuellement.

Voyons maintenant comment est produite la statistique de test.



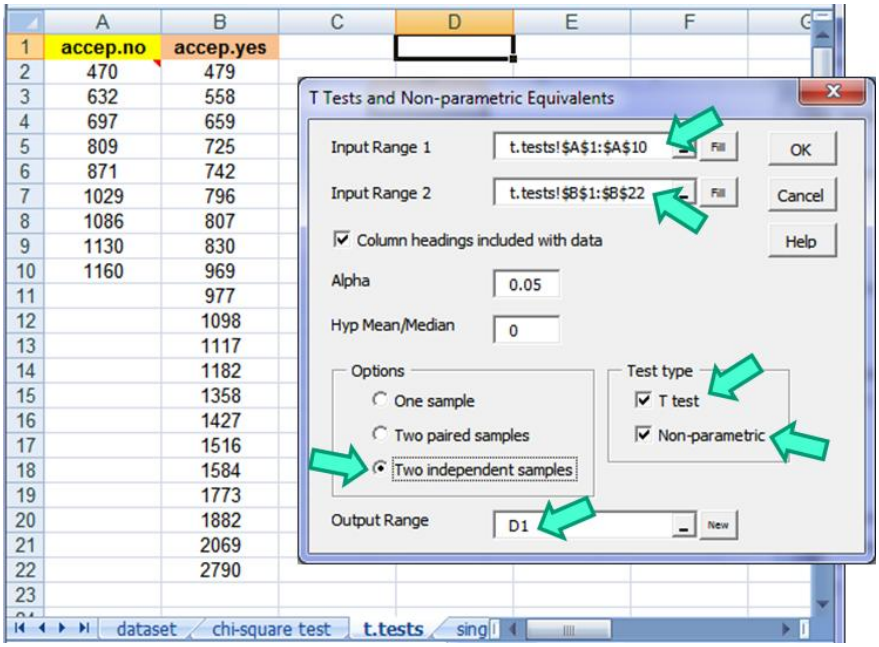
Le KHI-2 de Pearson est calculé à l'aide d'une fonction CHI_STAT2(.) propre à Real Statistics. Elle prend en entrée les effectifs observés (surlignés en orange) et théoriques (sous indépendance, surlignés en bleu). Ainsi, la libraire propose un grand nombre de nouvelles fonctions statistiques que nous pouvons utiliser directement, pour peu que l'on sache les paramétrer correctement. La liste des fonctions est accessible en ligne[10].

4.2 Test de comparaison de 2 moyennes (échantillons indépendants)

Pour illustrer le test de comparaison de moyennes de 2 échantillons indépendants, nous utilisons les variables INCOME.PER.HEAD et ACCEPTATION. Nous souhaitons savoir si, en moyenne, les revenus par tête sont identiques dans les deux groupes définies par l'acceptation c.-à-d. les personnes qui se voient accepter (vs. refuser) leur demander de crédit.

Nous copions les données dans la feuille « **t.tests** ». Nous les organisons de manière à pouvoir effectuer les calculs à l'aide de la procédure de Real Statistics. Nous créons pour cela 2 colonnes : les valeurs observées de income.per.head pour les personnes dont la demande a été refusée (acceptation = no), idem pour ceux qui ont été approuvés (acceptation = yes). Nous actionnons l'item « **T Tests and Non-parametric equivalents** » dans la fenêtre de lancement. Une boîte de paramétrage apparaît.

[10] Fonctions statistiques, http://www.real-statistics.com/excel-capabilities/supplemental-functions/ ; Analyse multivariée, http://www.real-statistics.com/excel-capabilities/real-statistics-multivariate-functions/; Traitement des données manquantes, http://www.real-statistics.com/excel-capabilities/real-statistics-advanced-missing-data-functions/



Nous indiquons : les 2 plages de données à traiter (Input Range), nous menons un test pour échantillons indépendants (Options : two independent samples), nous effectuons les tests paramétriques (Test type : T test) et non-paramétriques (Test type : Non-parametric), les sorties sont placées en D1 (output range).

SUMMARY. Cette section indique les statistiques descriptives conditionnelles (effectifs, moyennes, variances). Le D de Cohen correspond au rapport entre la différence des moyennes et l'écart-type commun. C'est une mesure normalisée, descriptive, permettant d'évaluer l'importance de l'écart entre les moyennes[11].

SUMMARY			Hyp Mean D	0
Groups	*Count*	*Mean*	*Variance*	*Cohen d*
accep.no	9	876.00	59201.00	
accep.yes	21	1206.57	331747.96	
Pooled			253877.40	0.656

T TEST : Equal Variances. Elle indique le résultat des tests de comparaison unilatéraux et bilatéraux sous l'hypothèse d'égalité des variances conditionnelles (T-TEST, onglet STATISTICS dans TANAGRA).

T TEST: Equal Variances				Alpha	0.05				
	std err	*t-stat*	*df*	*p-value*	*t-crit*	*lower*	*upper*	*sig*	*effect r*
One Tail	200.7436	1.6467	28	0.055396	1.7011			no	0.2971
Two Tail	200.7436	1.6467	28	0.110792	2.0484	-741.7761	80.6332	no	0.2971

T TEST : Unequal Variances. On s'affranchit de l'hypothèse d'homoscédasticité ici (T-TEST UNEQUAL VARIANCE dans Tanagra ; attention, la p-value est légèrement différente parce que

[11] http://en.wikipedia.org/wiki/Effect_size#Cohen.27s_d ; nous y reviendrons plus bas (section 4.3).

Tanagra utilise l'entier le plus proche pour les degrés de liberté fractionnaires, il est vraisemblable que Real Statistics s'appuie sur une interpolation linéaire).

T TEST: Unequal Variances				Alpha	0.05				
	std err	*t-stat*	*df*	*p-value*	*t-crit*	*lower*	*upper*	*sig*	*effect r*
One Tail	149.5841	2.2099	27.9906	0.017729	1.7011			yes	0.3854
Two Tail	149.5841	2.2099	27.9906	0.035457	2.0484	-636.9806	-24.1622	yes	0.3854

MANN-WHITNEY TEST for Two Independent Samples indique les résultats du test non paramétrique de comparaison de populations[12].

Real Statistics

Mann-Whitney Test for Two Independent Samples

	accep.no	accep.yes
count	9	21
median	871	1098
rank sum	109	356
U	125	64

	one tail	two tail
alpha	0.05	
U	64	
mean	94.50000	
std dev	22.09638	
z-score	-1.38032	
effect r	0.25201	
p-value	0.08374	0.16749
sig	no	no

Tanagra

Mann-Whitney U	64.00000
E(U)	94.50000
V(U)	488.25000
|Z|	1.38032
P(>|Z|)	0.16749

Les résultats[13] sont cohérents avec le composant MANN-WHITNEY COMPARISON de Tanagra. Ce dernier affiche uniquement la p-value du test bilatéral.

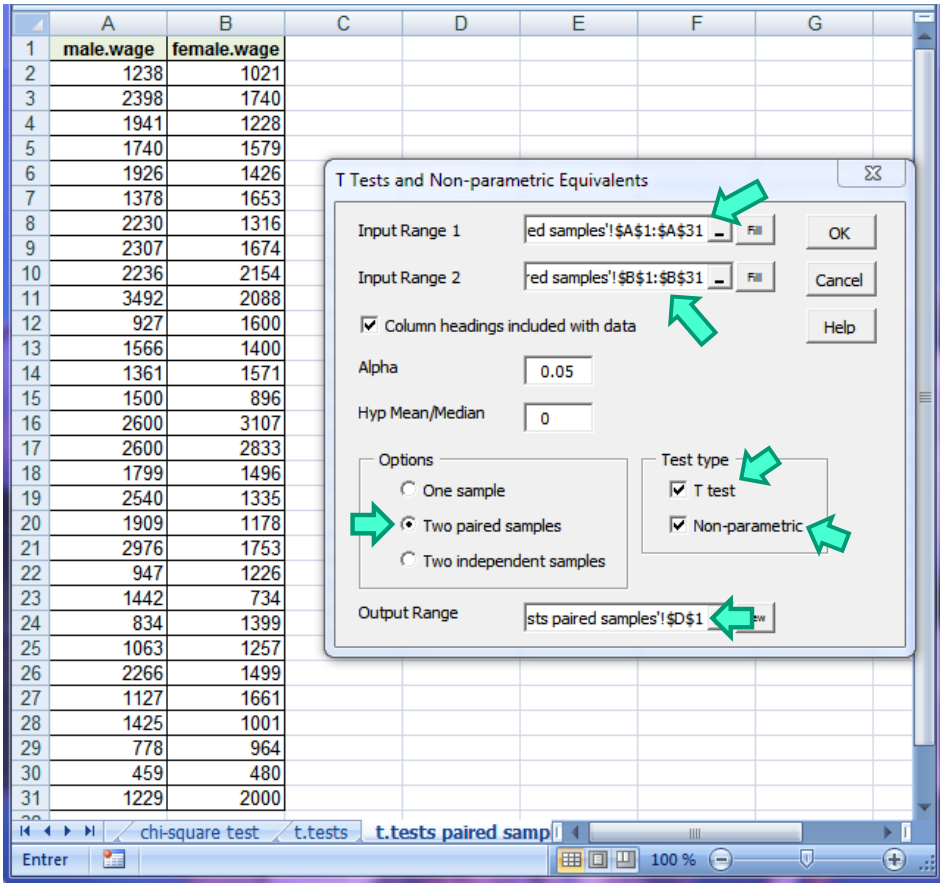
4.3 Test de comparaison (échantillons appariés)

Pour illustrer le test pour échantillons appariés, nous utilisons les variables MALE.WAGE et FEMALE.WAGE. L'objectif est de comparer les salaires à l'intérieur des ménages des demandeurs de crédit. On répond à la question : dans les couples, l'homme et la femme ont-ils – en moyenne – des niveaux de salaires différents ?

Nous copions les 2 colonnes dans la feuille « **t.tests paired samples** ». Nous sélectionnons l'item « **T Tests and Non-parametric equivalents** » dans la fenêtre de démarrage. Nous paramétrons la méthode comme suit.

[12] R. Rakotomalala, « Comparaison de populations – Tests non paramétriques », Août 2008.

[13] Le U critique est disponible à partir de la version 2.15 de Real Statistics.



Test paramétrique. La première partie des résultats concerne le test paramétrique. Real Statistics affiche les p-value pour les tests équilatéraux et bilatéraux.

T Test: Two Paired Samples

SUMMARY			Alpha	0.05		Hyp Mean Diff	0	
Groups	*Count*	*Mean*	*Std Dev*	*Std Err*	*t*	*df*	*Cohen d*	*Effect r*
male.wage	30	1741.1333	514328.1885					
female.wage	30	1508.9667	302040.0333					
Difference	30	232.1667	596.8388	108.9674	2.1306	29	0.3956	0.3679

T TEST	*p-value*	*t-crit*	*lower*	*upper*	*sig*
One Tail	0.0209	1.6991			yes
Two Tail	0.0417	2.0452	9.3034	455.0299	yes

Note : Taille d'effet (effect size). Arrêtons-nous un instant sur le tableau de résultat. J'avais remarqué que Real Statistics affichait systématiquement le d de Cohen et la taille d'effet corrélation r. Pourquoi ? C'est une pratique peu usuelle dans les ouvrages francophones. La taille d'effet mesure l'intensité d'un phénomène (relation entre 2 variables, différences entre 2 valeurs estimées). Il s'agit d'un indicateur et non pas une statistique inférentielle permettant de conclure ou pas à l'existence du phénomène dans la population. Je note surtout qu'elle annihile le rôle de la taille de l'échantillon dans les calculs. Et on comprend pourquoi. On

reproche souvent à la statistique inférentielle de produire des résultats systématiquement significatifs (rejet de l'hypothèse nulle) dès que la taille de l'échantillon augmente un tant soit peu. Avec une mesure normalisée, nous évitons cet écueil. Mais elle est purement descriptive. Concernant le *d* de Cohen par exemple, on admet généralement que l'effet est faible autour de 0.2, moyen autour de 0.5, fort autour de 0.8. Mais ce sont des repères très grossiers, tout dépend du domaine étudié[14,15]. Il n'en reste pas moins que ces mesures sont en relation directe avec la statistique de test **t**, mais déflatée de la taille de l'échantillon n = 30. En posant df = n - 1 = 29, les degrés de libertés, les formules utilisées ici s'écrivent (elles sont visibles dans les cellules contenant les résultats) :

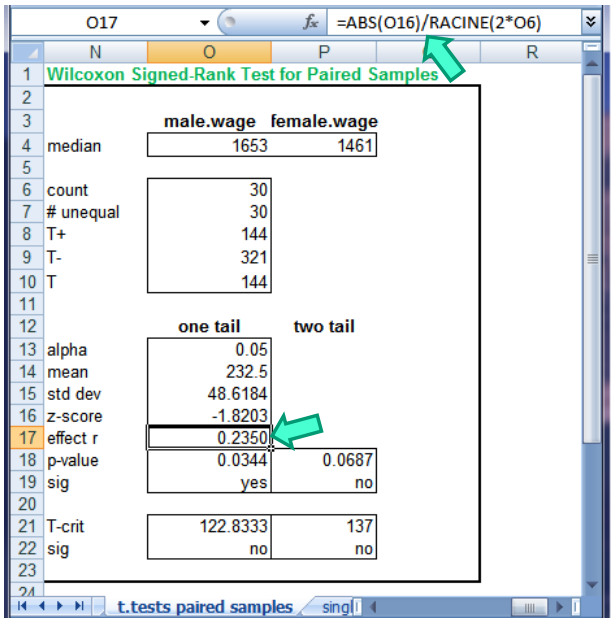
$$d = \frac{|t|}{\sqrt{df}} = \frac{2.1306}{\sqrt{29}} = 0.3956$$

Et

$$r_p = \sqrt{\frac{t^2}{t^2 + df}} = \sqrt{\frac{2.1306^2}{2.1306^2 + 29}} = 0.3679$$

L'écart des salaires est modéré, même il s'avère significatif à 5% sur un échantillon de n = 30 observations avec un p-value = 0.0417.

Test non-paramétrique. Real Statistics exploite le test des rangs signés de Wilcoxon.



On remarquera que la taille d'effet « r » s'écrit différemment dans ce cas :

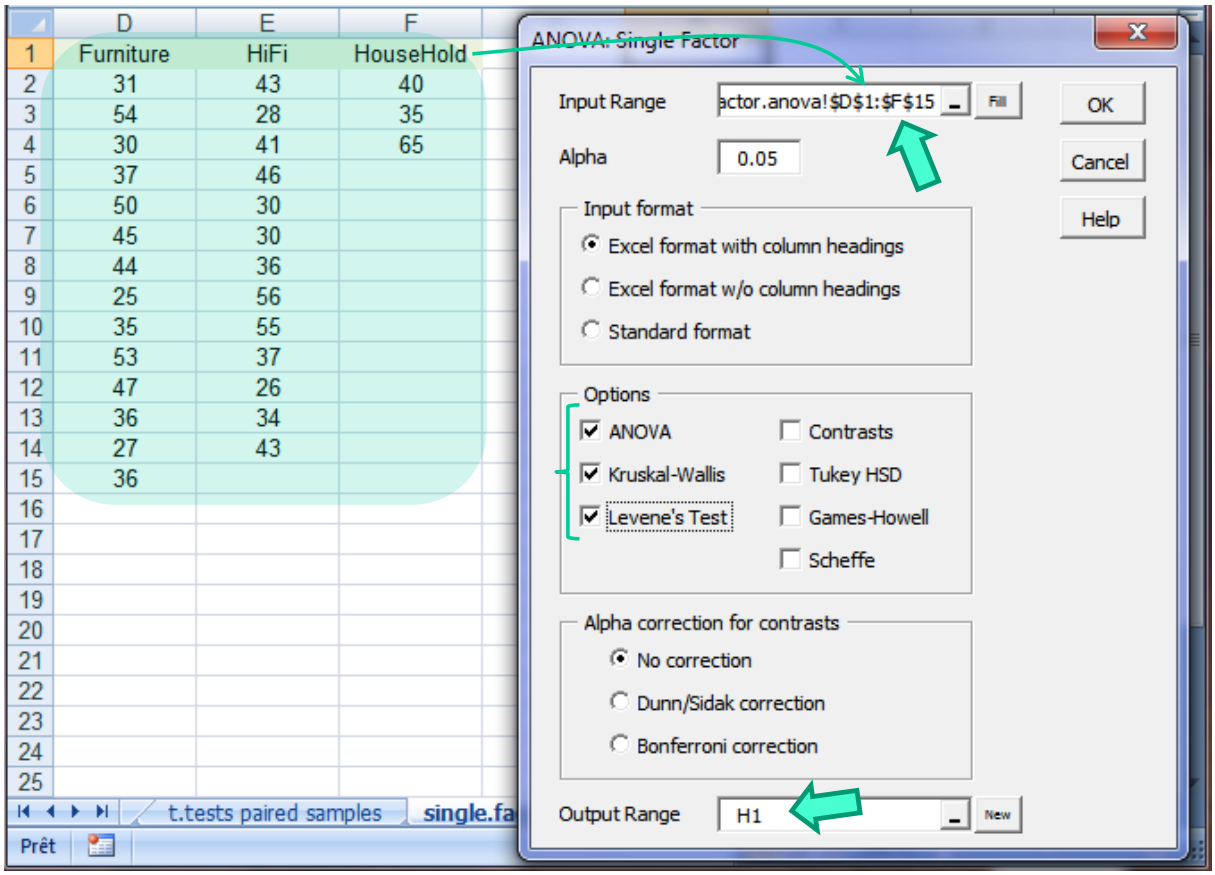
$$r_w = \frac{|z|}{\sqrt{2 \times n}} = \frac{1.8203}{\sqrt{2 \times 30}} = 0.2350$$

[14] http://www.tea.state.tx.us/Best_Practice_Standards/How_To_Interpret_Effect_Sizes.aspx

[15] http://www.leeds.ac.uk/educol/documents/00002182.htm (la présentation est particulièrement étayée).

4.4 Analyse de variance (Anova) à 1 facteur

En schématisant, on peut considérer l'ANOVA à 1 facteur comme une généralisation du test de comparaison de moyennes pour (K > 2) échantillons indépendants. Nous cherchons à savoir si l'âge moyen des personnes est différent selon le type d'achat motivant la demande de crédit. Nous copions les variables REASON et AGE dans la feuille « **single.factor.anova** ». Nous réorganisons les données en identifiant liste des valeurs d'AGE pour chaque modalité de REASON. Nous activons alors l'item « Single Factor Anova » dans la fenêtre de démarrage et nous paramétrons la méthode comme suit.



Qu'importe si certaines cellules sont vides (les effectifs conditionnels ne sont pas forcément identiques). L'outil s'adapte automatiquement. Nous demandons en plus le test non-paramétrique de Kruskal-Wallis, et le test de comparaison des variances conditionnelles de Levene. Nous laissons de côté en revanche les comparaisons multiples de moyennes (Contrasts, Tukey HSD, etc.).

Analyse de variance à 1 facteur. Nous disposons du tableau des statistiques conditionnelles (DESCRIPTION) (moyenne, variance, etc.), puis celui de la décomposition de la variance incluant la statistique de test F et la probabilité critique (p-value).

ANOVA: Single Factor

DESCRIPTION Alpha 0.05

Groups	*Count*	*Sum*	*Mean*	*Variance*	*SS*	*Std Err*	*Lower*	*Upper*
Furniture	14	550	39.2857	91.4505	1188.8571	2.7339	33.3795	45.1919
HiFi	13	505	38.8462	93.3077	1119.6923	2.8371	32.6647	45.0276
HouseHold	3	140	46.6667	258.3333	516.6667	5.9059	21.2558	72.0775

ANOVA

Sources	*SS*	*df*	*MS*	*F*	*P value*	*F crit*	*RMSSE*	*Omega Sq*
Between Groups	156.9505	2	78.4753	**0.749972**	**0.481966**	3.3541	0.4295	-0.0170
Within Groups	2825.2161	27	104.6376					
Total	2982.1667	29	102.8333					

Test de Kruskal-Wallis. Il s'agit du pendant non paramétrique de l'ANOVA à 1 facteur.

Real Statistics

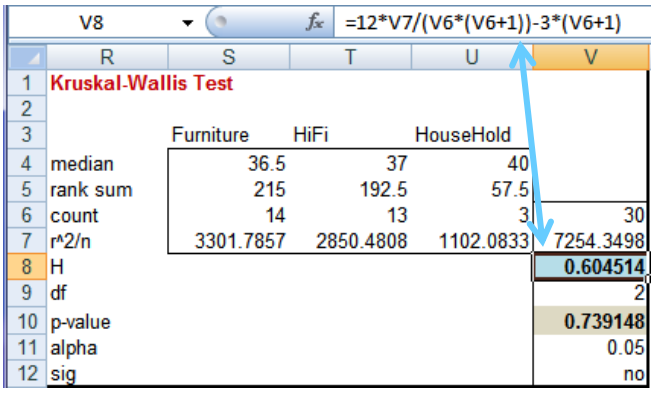
Kruskal-Wallis Test

	Furniture	HiFi	HouseHold	
median	36.5	37	40	
rank sum	215	192.5	57.5	
count	14	13	3	30
r^2/n	3301.7857	2850.4808	1102.0833	7254.3498
H				**0.604514**
df				2
p-value				**0.739148**
alpha				0.05
sig				no

Tanagra

Description					Statistical test		
Value	**Examples**	**Average**	**Rank sum**	**Rank mean**	**Statistics**	**Value**	**Proba**
Furniture	14	39.2857	215	15.3571	Kruskal-Wallis	0.604514	0.739148
HiFi	13	38.8462	192.5	14.8077	KW (corr.ties)	0.605997	0.738600
HouseHold	3	46.6667	57.5	19.1667			
All	30	39.8333	465	15.5			

Real Statistics utilise bien les rangs moyens pour les ex-aequos (les sommes des rangs sont identiques à ceux de Tanagra). Il fournit en revanche la statistique non corrigée H = 0.604514. Nous distinguons la formule utilisée dans la cellule V8.



Effectivement, elle ne comporte pas la correction pour ex-aequos[16].

Test de Levene. Il est destiné à vérifier l'égalité des variances conditionnelles. Ce test est autrement plus robuste que celui de Bartlett. 3 variantes sont proposées : celle basée sur la moyenne dans les groupes, sur la médiane, et sur la moyenne tronquée (trimmed mean)[17].

[16] Voir http://en.wikipedia.org/wiki/Kruskal-Wallis_one-way_analysis_of_variance

[17] http://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm

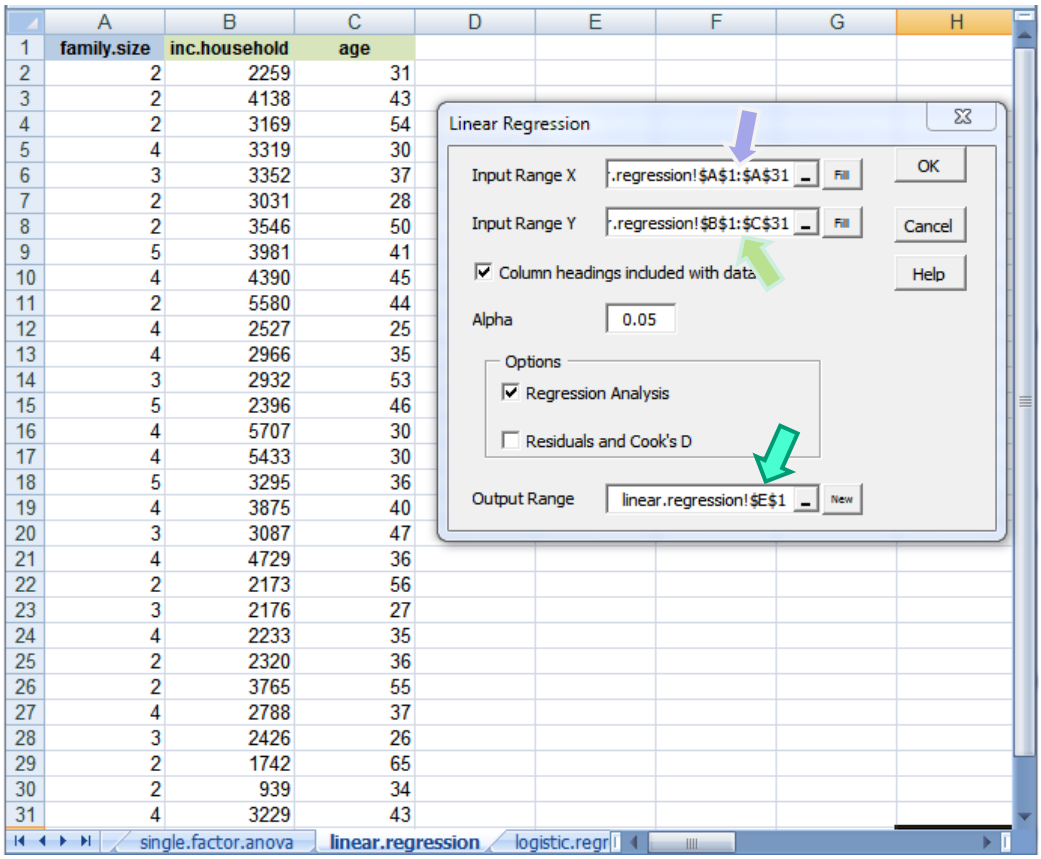
Levene's Tests	
type	p-value
means	0.3731
medians	0.8573
trimmed	0.3731

Les observations sont compatibles avec l'hypothèse d'égalité des variances au risque 5%.

4.5 Régression linéaire multiple

On cherche à expliquer la taille de la famille en fonction du revenu du ménage et de l'âge du demandeur de crédit pour illustrer la régression. Je suis d'accord, ça n'a pas vraiment de sens. L'objectif est simplement de détailler les sorties de Real Statistics.

Nous copions les 3 colonnes - dans l'ordre FAMILY.SIZE, INC.HOUSEHOLD, AGE - dans la feuille « **linear.regression** ». Nous sélectionnons « **Linear Regression** » dans la fenêtre de démarrage. Nous spécifions les paramètres suivants.



	A	B	C
1	family.size	inc.household	age
2	2	2259	31
3	2	4138	43
4	2	3169	54
5	4	3319	30
6	3	3352	37
7	2	3031	28
8	2	3546	50
9	5	3981	41
10	4	4390	45
11	2	5580	44
12	4	2527	25
13	4	2966	35
14	3	2932	53
15	5	2396	46
16	4	5707	30
17	4	5433	30
18	5	3295	36
19	4	3875	40
20	3	3087	47
21	4	4729	36
22	2	2173	56
23	3	2176	27
24	4	2233	35
25	2	2320	36
26	2	3765	55
27	4	2788	37
28	3	2426	26
29	2	1742	65
30	2	939	34
31	4	3229	43

Les colonnes des variables explicatives doivent être contigües, comme pour la fonction DROITEREG d'Excel. L'option « Residuals and Cook's D », non cochée dans notre exemple, produit les indicateurs permettant de détecter les points atypiques et/ou influents (levier, résidus studentisés, distance de Cook, etc.). Attention, la taille du tableau est conséquente si le nombre d'observations est élevé.

Avec les options que nous avons sélectionnées, Real Statistics fournit un diagnostic global de la régression (R, R^2 ajusté, écart type estimé de l'erreur, etc.), le tableau d'analyse de variance, et le tableau des coefficients incluant les tests de significativité et les intervalles de confiance.

Regression Analysis

OVERALL FIT

Multiple R	0.3979
R Square	0.1583
Adjusted R Square	0.0960
Standard Error	1.0112
Observations	30

ANOVA Alpha 0.05

	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	2	5.1934	2.5967	2.5396	0.0976	no
Residual	27	27.6066	1.0225			
Total	29	32.8				

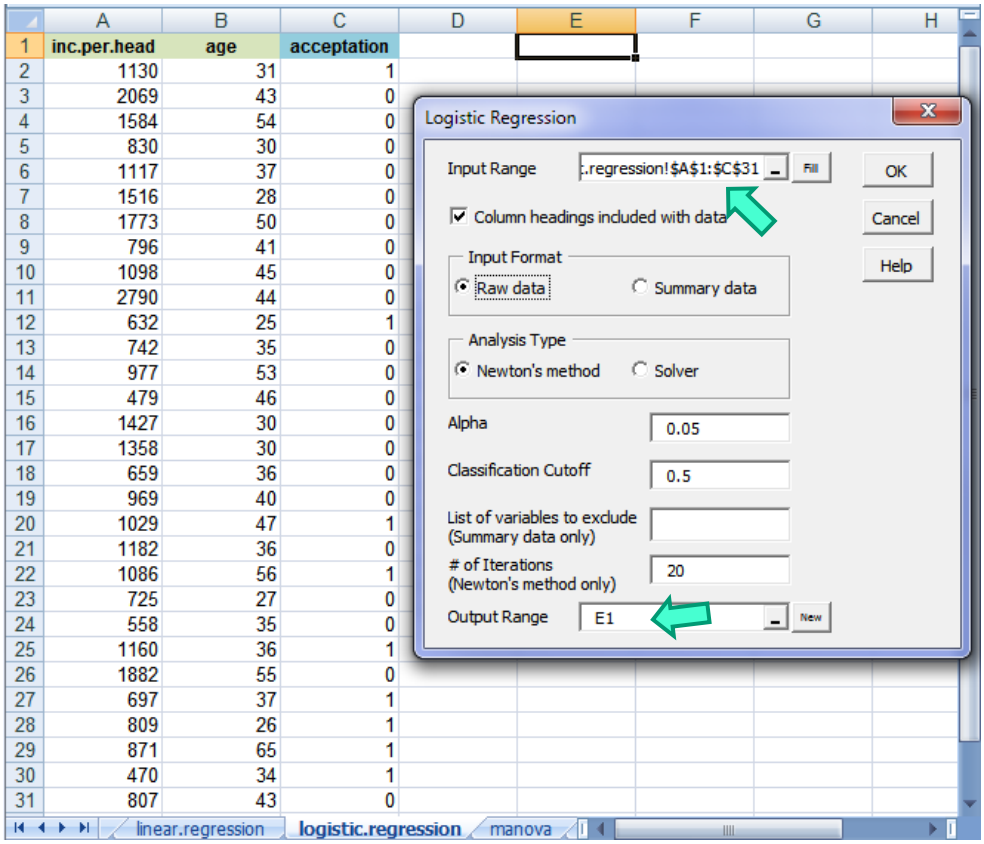
	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*
Intercept	3.8423	0.9562	4.0181	0.0004	1.8802	5.8043
inc.household	0.0002	0.0002	1.2649	0.2167	-0.0001	0.0006
age	-0.0333	0.0185	-1.7959	0.0837	-0.0713	0.0047

Mine de rien, notre régression n'est pas si désastreuse que cela. Elle est globalement significative à 10%. Pour un si faible effectif (n = 30), ce n'est pas anodin. Après, interpréter les résultats est une autre histoire, je ne m'y risquerai pas.

4.6 Régression logistique

Avec la régression logistique, nous cherchons à expliquer les valeurs prises par une variable dépendante qualitative binaire (ACCEPTATION) à partir de variables indépendantes quantitatives (INC.PER.HEAD et AGE). Un recodage préalable de ACCEPTATION est nécessaire, nous posons 1 lorsque le crédit est refusé (acceptation = no), 0 dans le cas contraire. Nous avons fait ce choix parce que nous souhaitons mettre en évidence les mobiles du refus d'une demande de crédit.

Les données préparées sont copiées dans la feuille « **logistic.regression** », nous actionnons l'item « **Logistic Regression** » dans la fenêtre de démarrage.



	A	B	C
1	inc.per.head	age	acceptation
2	1130	31	1
3	2069	43	0
4	1584	54	0
5	830	30	0
6	1117	37	0
7	1516	28	0
8	1773	50	0
9	796	41	0
10	1098	45	0
11	2790	44	0
12	632	25	1
13	742	35	0
14	977	53	0
15	479	46	0
16	1427	30	0
17	1358	30	0
18	659	36	0
19	969	40	0
20	1029	47	1
21	1182	36	0
22	1086	56	1
23	725	27	0
24	558	35	0
25	1160	36	1
26	1882	55	0
27	697	37	1
28	809	26	1
29	871	65	1
30	470	34	1
31	807	43	0

INPUT RANGE désigne la plage de cellules des données, sans distinction du rôle des variables. Pour que la procédure fonctionne, la variable cible doit être située en dernière colonne (la plus à droite), et codée 0/1 (laisser les valeurs yes/no fait échouer la procédure).

Les résultats sont touffus et disséminés à plusieurs endroits. Essayons d'y voir plus clair.

LL0	-18.3259
LL1	-16.5698
Chi-Sq	3.5123
df	2
p-value	0.1727
alpha	0.05
sig	no
R-Sq (L)	0.0958
R-Sq (CS)	0.1105
R-Sq (N)	0.1566
Hosmer	27.0050
df	28
p-value	0.5180
alpha	0.05
sig	no

Evaluation globale de la régression. Ce tableau regroupe les résultats globaux de la régression. Nous observons, entres autres, la log-vraisemblance du modèle (LL1 = -16.5698), la log-vraisemblance du modèle trivial réduit à la constante (LL0 = -18.3259). A partir de ces informations, Real Statistics calcule la statistique du test de pertinence globale (Chi-Sq). La régression n'est pas significative à 5% avec une p-value de 0.1727. Différentes valeurs de pseudo-R2 sont proposées (McFadden, Cox and Snell, Nagelkerke). Le test de Hosmer Lemeshow sert à confronter les scores observés et prédits. La « p-value » est égale à 0.5180, le modèle est compatible avec les données[18].

[18] Ricco Rakotomalala, « Pratique de la Régression Logistique – Régression Logistique Binaire et Polytomique », 2014.

Matrice de confusion. La CLASSICATION TABLE confronte les valeurs observées et prédites de la variable dépendante ACCEPTATION. « Accuracy » correspond en réalité à la sensibilité. Par ex., il y a 9 « acceptation = no » observées (Suc-Obs), 1 a été classé correctement, 8 a été attribuée à l'autre classe (acceptation = yes, Fail-Pred). La sensibilité est de donc de 0.111. Le taux de succès (1 – taux d'erreur) du modèle est de 0.7 (surlignée en brun).

Classification Table

	Suc-Obs	Fail-Obs	
Suc-Pred	1	1	2
Fail-Pred	8	20	28
	9	21	30
Accuracy	0.111	0.952	0.7
Cutoff	0.5		

Nous pouvons modifier interactivement le seuil d'affectation (CUTOFF, surlignée en bleu), la matrice de confusion est automatiquement remise à jour. Par ex., pour améliorer la sensibilité du modèle, nous pouvons abaisser cette valeur seuil à CUTOFF = 0.3. La sensibilité s'améliore (0.778), mais au détriment de la performance globale (taux de succès = 0.6).

Classification Table

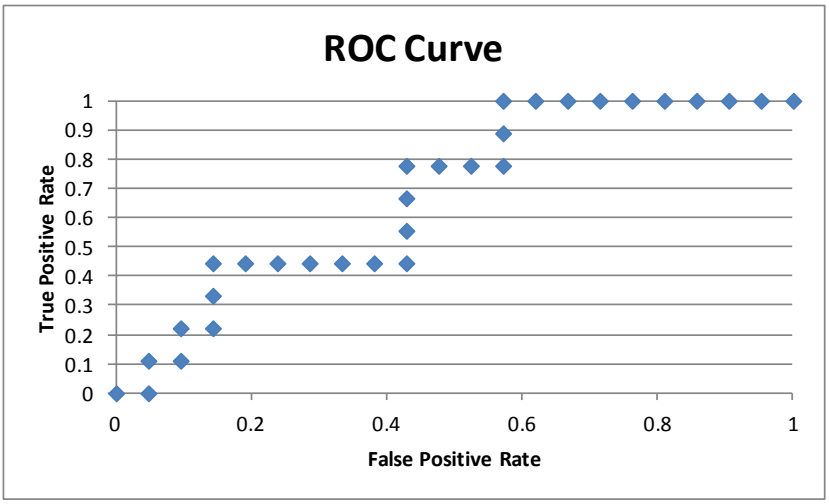
	Suc-Obs	Fail-Obs	
Suc-Pred	7	10	17
Fail-Pred	2	11	13
	9	21	30
Accuracy	0.778	0.524	0.6
Cutoff	0.3		

Coefficients estimés. Le tableau des coefficients estimés inclut leurs écarts-type, les statistiques de Wald, les p-value associées, les odds-ratio [exp(coefficient)], et leurs intervalles de confiance.

	coeff b	*s.e.*	*Wald*	*p-value*	*exp(b)*	*lower*	*upper*
Intercept	0.4389	1.8613	0.0556	0.8136	1.5510		
inc.per.head	-0.0020	0.0013	2.4925	0.1144	0.9980	0.9955	1.0005
age	0.0188	0.0429	0.1916	0.6616	1.0190	0.9368	1.1083

Ni le revenu par tête (inc.per.head), ni l'âge, ne semblent influer sur le rejet des demandes. Le modèle n'étant pas globalement significatif, on pouvait s'attendre à ce résultat.

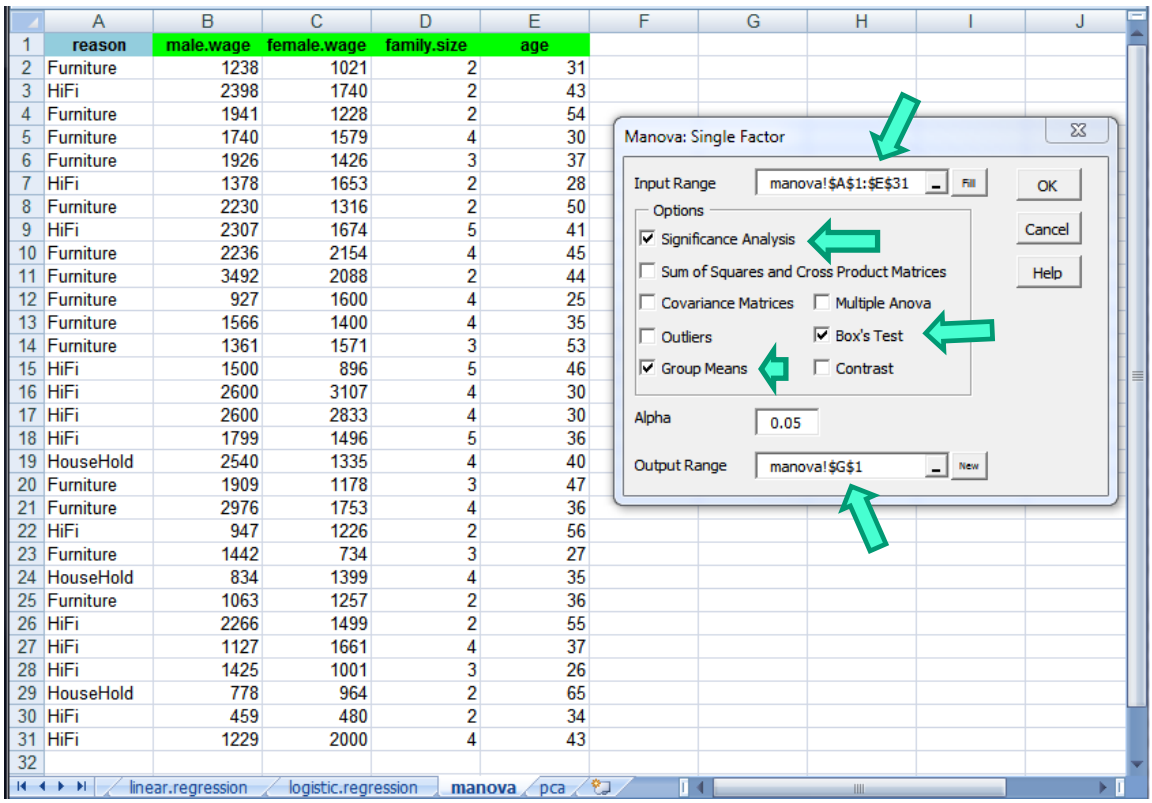
Courbe ROC. Real Statistics produit d'autres tableaux, l'une destinée au calcul de la statistique de Hosmer et Lemeshow, l'autre à la courbe ROC, laquelle est automatiquement dessinée dans un graphique « nuage de points ».



4.7 MANOVA

La MANOVA (multivariate analysis of variance) est une généralisation multivariée de l'ANOVA. On cherche à percevoir les différences entre les groupes, en prenant en compte le rôle simultané de plusieurs variables. Dans notre exemple, nous cherchons à savoir si les caractéristiques des personnes (MALE.WAGE, FEMALE.WAGE, FAMILY.SIZE, AGE) sont différentes selon le type d'achat motivant la demande de crédit (REASON).

Nous copions les variables dans une nouvelle feuille « **manova** ». Nous actionnons « **Single Factor Manova** » dans la fenêtre de démarrage.



	reason	male.wage	female.wage	family.size	age
2	Furniture	1238	1021	2	31
3	HiFi	2398	1740	2	43
4	Furniture	1941	1228	2	54
5	Furniture	1740	1579	4	30
6	Furniture	1926	1426	3	37
7	HiFi	1378	1653	2	28
8	Furniture	2230	1316	2	50
9	HiFi	2307	1674	5	41
10	Furniture	2236	2154	4	45
11	Furniture	3492	2088	2	44
12	Furniture	927	1600	4	25
13	Furniture	1566	1400	4	35
14	Furniture	1361	1571	3	53
15	HiFi	1500	896	5	46
16	HiFi	2600	3107	4	30
17	HiFi	2600	2833	4	30
18	HiFi	1799	1496	5	36
19	HouseHold	2540	1335	4	40
20	Furniture	1909	1178	3	47
21	Furniture	2976	1753	4	36
22	HiFi	947	1226	2	56
23	Furniture	1442	734	3	27
24	HouseHold	834	1399	4	35
25	Furniture	1063	1257	2	36
26	HiFi	2266	1499	2	55
27	HiFi	1127	1661	4	37
28	HiFi	1425	1001	3	26
29	HouseHold	778	964	2	65
30	HiFi	459	480	2	34
31	HiFi	1229	2000	4	43

Nous spécifions toute la plage de cellules dans INPUT RANGE. La variable définissant les groupes doit être située en première colonne (la plus à gauche). Selon les options sélectionnées, nous obtenons plusieurs blocs de résultats.

Group Means indique les moyennes conditionnelles.

Group Means

	male.wage	female.wage	family.size	age	Count
Furniture	1860.50	1450.36	3.00	39.29	14
HiFi	1695.00	1635.85	3.38	38.85	13
HouseHold	1384.00	1232.67	3.33	46.67	3
Total	1741.13	1508.97	3.20	39.83	30

MANOVA fournit les tests de significativité globale. Plusieurs statistiques sont proposées.

MANOVA

	stat	*F*	*df1*	*df2*	*p-value*	*eta-sq*
Pillai Trace	0.22396	0.78815	8	50	0.61530	0.11198
Wilk's Lambda	0.78787	0.75965	8	48	0.63934	0.11238
Hotelling Trace	0.25423	0.73091	8	46	0.66362	0.11278
Roy's Lg Root	0.16087					

Box's test diagnostique l'égalité des dispersions conditionnelles.

Real Statistics

Box's Test

M	235.2691
F	7.1465
df1	20
df2	281.0313
p-value	0.0000

Tanagra

Tests results

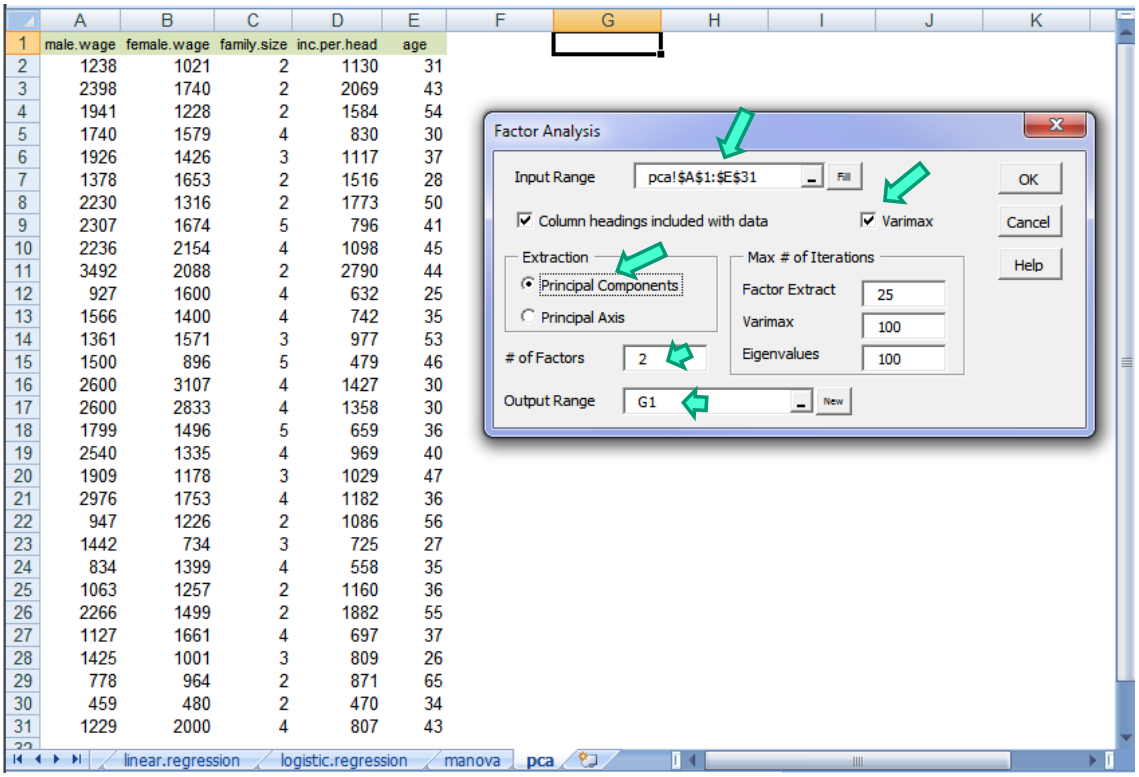
Stat	Value	p-value
T [CHI-2 (20)]	42.7890	0.0022

Tanagra s'appuie sur l'approximation du KHI-2, Real Statistics sur celle de Fisher[19]. Cette dernière est plus appropriée sur un effectif aussi faible.

4.8 Analyse en composantes principales (ACP)

Pour illustrer l'ACP, nous utilisons la plupart des variables quantitatives, à savoir : MALE.WAGE, FEMALE.WAGE, FAMILY.SIZE, INC.PER.HEAD et AGE. Nous les copions dans la feuille « **pca** » et nous actions l'item « **Factor Analysis** » dans la fenêtre de démarrage. Nous demandons la construction des 2 premiers facteurs, avec une rotation VARIMAX.

[19] http://www.real-statistics.com/multivariate-statistics/boxs-test-equality-covariance-matrices/boxs-test-basic-concepts/



Les résultats sont décomposés en plusieurs sections.

Descriptive Statistics. Moyenne, écart-type, asymétrie et aplatissement.

Descriptive statistics

	male.wage	female.wage	family.size	inc.per.head	age
Mean	1741.133	1508.967	3.200	1107.400	39.833
Std dev	717.167	549.582	1.064	518.520	10.141
Skewness	0.413	1.049	0.125	1.450	0.597
Kurtosis	-0.238	2.155	-1.387	2.605	-0.218

Correlation matrix. La matrice des corrélations.

Correlation Matrix

	male.wage	female.wage	family.size	inc.per.head	age
male.wage	1	0.58374	0.13338	0.67598	0.03897
female.wage	0.58374	1	0.32019	0.39480	-0.15946
family.size	0.13338	0.32019	1	-0.55074	-0.32933
inc.per.head	0.67598	0.39480	-0.55074	1	0.26341
age	0.03897	-0.15946	-0.32933	0.26341	1

Inverse of Correlation Matrix. L'inverse de la matrice des corrélations.

Inverse of Correlation Matrix

	male.wage	female.wage	family.size	inc.per.head	age
male.wage	6.14724	1.10400	-5.26376	-7.54008	0.18910
female.wage	1.10400	2.62204	-2.72757	-3.38039	0.36725
family.size	-5.26376	-2.72757	7.33245	8.70155	-0.10709
inc.per.head	-7.54008	-3.38039	8.70155	12.39351	-0.64410
age	0.18910	0.36725	-0.10709	-0.64410	1.18559

Elle servira surtout à calculer la matrice des corrélations partielles qui suit.

Partial Correlation Matrix. Elle indique la liaison nette entre les variables, en retranchant l'influence de toutes les autres.

Partial Correlation Matrix

	male.wage	female.wage	family.size	inc.per.head	age
male.wage	1	-0.2750	0.7840	0.8639	-0.0700
female.wage	-0.2750	1	0.6221	0.5930	-0.2083
family.size	0.7840	0.6221	1	-0.9128	0.0363
inc.per.head	0.8639	0.5930	-0.9128	1	0.1680
age	-0.0700	-0.2083	0.0363	0.1680	1

KMO. L'indice KMO (Kaiser – Mayer – Olkin, connu également sous l'appellation MSA, measure of sampling adequacy) indique le degré de compressibilité des données c.-à-d. la redondance des variables, et la possibilité de la (cette redondance) prendre en compte dans l'ACP[20]. En rouge, nous avons l'indice KMO global.

KMO

male.wage	female.wage	family.size	inc.per.head	age	
0.36176	0.42139	0.22467	0.33466	0.72457	**0.33889**

Eigenvalues and eigenvectors. Les valeurs propres sont situées sur la première ligne (en bleu) ; les vecteurs propres sont situés en dessous (en vert), elles sont organisées en colonnes c.-à-d. la 1ère colonne correspond au 1er vecteur propre, etc. Real Statistics utilise une fonction dédiée **eVectors(.)** pour les produire.

Eigenvalues and eigenvectors

2.13244	1.71007	0.74148	0.37538	0.04063
0.59878	0.20713	0.10561	-0.62300	0.44640
0.46557	0.43743	0.05546	0.74213	0.19513
-0.14051	0.67091	0.46648	-0.20546	-0.51993
0.62071	-0.26395	-0.22910	-0.03236	-0.70108
0.14029	-0.49595	0.84598	0.13365	0.02831

Full load matrix. Cette matrice correspond aux corrélations des variables avec les composantes principales.

Full Load Matrix

	1	2	3	4	5
male.wage	0.87440	0.27086	0.09094	-0.38170	0.08998
female.wage	0.67987	0.57202	0.04776	0.45469	0.03933
family.size	-0.20518	0.87734	0.40168	-0.12588	-0.10480
inc.per.head	0.90641	-0.34517	-0.19727	-0.01983	-0.14131
age	0.20487	-0.64855	0.72847	0.08189	0.00571

[20] « ACP sous R – Indice KMO et test de Barteltt », Mai 2012.

Le carré des valeurs correspond au cosinus carré (COS²), leur somme en ligne vaut 1 forcément.

Scree plot. Ce tableau fournit les valeurs propres et les proportions d'inertie restituées par les axes. Real Statistics lui associe l'éboulis des valeurs propres (scree plot).

Scree Plot

eValue	%	Cum %
2.1324	42.65%	42.65%
1.7101	34.20%	76.85%
0.7415	14.83%	91.68%
0.3754	7.51%	99.19%
0.0406	0.81%	100.00%

Factor matrix. Ces matrices représentent aux corrélations des variables avec les axes, avant (unrotated) et après (rotated) la rotation varimax. « Commun » correspond aux « communalities », la part de variance de variable traduite par les facteurs sélectionnés ; « specif » = 1 – « commun », la part d'information des variables non prise en compte par les facteurs sélectionnés.

Factor Matrix (unrotated)

	1	2	Commun	Specific
male.wage	0.8744	0.2709	0.8379	0.1621
female.wage	0.6799	0.5720	0.7894	0.2106
family.size	-0.2052	0.8773	0.8118	0.1882
inc.per.head	0.9064	-0.3452	0.9407	0.0593
age	0.2049	-0.6486	0.4626	0.5374
	2.1324	1.7101	3.8425	1.1575

Factor Matrix (rotated Varimax)

	1	2	Commun	Specific
male.wage	0.9149	-0.0304	0.8379	0.1621
female.wage	0.8297	0.3178	0.7894	0.2106
family.size	0.0935	0.8962	0.8118	0.1882
inc.per.head	0.7434	-0.6230	0.9407	0.0593
age	-0.0188	-0.6799	0.4626	0.5374
	2.0871	1.7554	3.8425	1.1575

Reproduced correlation matrix et Error matrix. La première représente l'information (les corrélations) reproduite sur les axes sélectionnés. Nous avons les « communalities » sur la diagonale principale. La seconde confronte la matrice des corrélations originelle avec la matrice estimée. Elle indique la fidélité de la représentation.

Reproduced Correlation Matrix

	male.wage	female.wage	family.size	inc.per.head	age
male.wage	0.83794	0.74941	0.05823	0.69907	0.00347
female.wage	0.74941	0.78943	0.36237	0.41880	-0.23171
family.size	0.05823	0.36237	0.81182	-0.48880	-0.61104
inc.per.head	0.69907	0.41880	-0.48880	0.94072	0.40955
age	0.00347	-0.23171	-0.61104	0.40955	0.46259

Error Matrix

	male.wage	female.wage	family.size	inc.per.head	age
male.wage	0.16206	-0.16567	0.07515	-0.02309	0.03551
female.wage	-0.16567	0.21057	-0.04218	-0.02399	0.07225
family.size	0.07515	-0.04218	0.18818	-0.06194	0.28171
inc.per.head	-0.02309	-0.02399	-0.06194	0.05928	-0.14614
age	0.03551	0.07225	0.28171	-0.14614	0.53741

On notera par exemple que l'information véhiculée par la variable AGE est mal représentée sur les 2 premiers facteurs.

Factor Scores. Ces coefficients permettent de calculer les coordonnées factorielles des individus à partir des variables originelles. Plusieurs formulations sont proposées : Regression Method, Bartlett's Method, Anderson-Rubin's Method[21].

Factor Scores Matrix - Regression Method

	1	2
male.wage	0.43931	0.01537
female.wage	0.41079	0.21165
family.size	0.07711	0.51626
inc.per.head	0.33552	-0.32992
age	-0.03343	-0.38980

Factor Scores Matrix - Bartlett's Method

	1	2
male.wage	0.38061	0.18059
female.wage	0.32362	0.29080
family.size	0.20232	0.49583
inc.per.head	0.48895	-0.61239
age	-0.04710	-0.12829

Factor Scores - Anderson-Rubin's Method

	360.55181	-194.01169
	-194.01169	197.94014

	1	2
male.wage	0.37899	0.17772
female.wage	0.32148	0.28381
family.size	0.19927	0.48105
inc.per.head	0.49150	-0.58808
age	-0.04632	-0.12443

[21] http://www.real-statistics.com/multivariate-statistics/factor-analysis/factor-scores/

5 Conclusion

L'add-in Real Statistics pour Excel est un travail remarquable à plusieurs égards. D'un point de vue fonctionnel, il permet de mener des études réelles. La manipulation est très simple. Les différentes sections des sorties sont parfaitement identifiées. Les calculs sont précis, du moins en ce qui concerne les procédures que j'ai pu tester. Mais le plus intéressant à mon sens est la documentation disponible sur le site de l'auteur (http://www.real-statistics.com/). Les méthodes sont parfaitement décrites, avec force exemples sur des petits jeux de données. Les fonctions spécifiques sont énumérées. Il est possible de les appeler directement dans notre classeur sans passer par l'interface dédiée de l'add-in. J'ai réellement eu beaucoup de plaisir à découvrir cette librairie.

Remarque : L'add-in est constamment mis à jour, j'invite les utilisateurs à consulter régulièrement le site web pour suivre les dernières améliorations. L'auteur m'a récemment signalé (version 2.15) l'introduction d'outils dédiés aux calculs de puissance statistique des tests.