

Objectif

TANAGRA intègre une nouvelle palette de composants dédiée aux statistiques non-paramétriques. Dans ce didacticiel, nous effectuons un tour d'horizon des méthodes implémentées à partir d'un fichier de données fictif représentant les caractéristiques des ménages selon leur revenu et leur habitation.

Les techniques reprises dans TANAGRA sont en très grande partie tirées du livre de Siegel et Castellan (**Sidney SIEGEL et John CASTELLAN, « Nonparametric Statistics for the Behavioral Sciences », McGraw-Hill, 1988**). J'en conseille la lecture à tous ceux qui s'intéressent de près ou de loin aux méthodes non-paramétriques en statistique.

Fichier

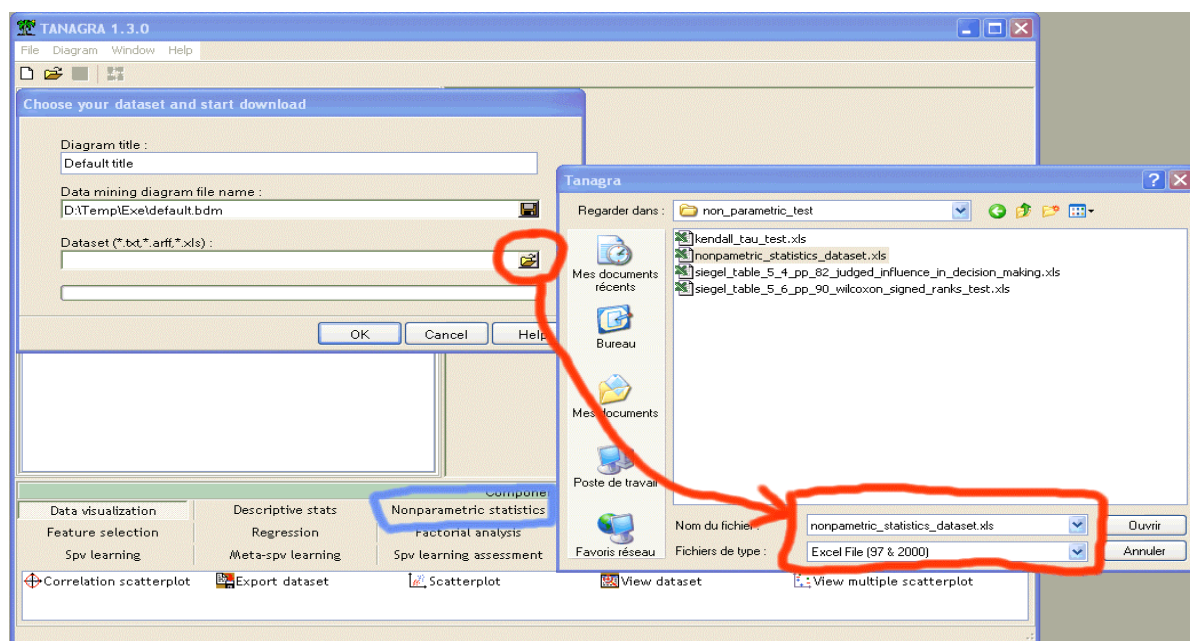
Notre fichier (NONPARAMETRIC_STATISTICS_DATASET.XLS) décrit 300 ménages à l'aide de 5 variables (salaire de l'homme, salaire de la femme, revenu du ménage [l'addition des deux salaires], l'habitation et le fait disposer d'un jardin ou pas).

Les exemples sont plus ou moins loufoques, le plus important est de montrer le mode de fonctionnement du logiciel et la lecture des résultats.

Statistiques non-paramétriques

Charger les données

Première étape toujours, charger le fichier de données dans un nouveau diagramme (FILE / NEW).



Comparaison de 2 populations

Nous voulons comparer le revenu total des ménages (INCOME) selon qu'ils occupent une habitation avec jardin ou pas (GARDEN – 2 modalités).

Placez le composant DEFINE STATUS dans le diagramme, mettez INCOME en TARGET et GARDEN en INPUT. Plusieurs méthodes peuvent être mises en œuvre, nous privilégions les méthodes qui s'appliquent uniquement à la comparaison de 2 populations dans cette section, il s'agit des méthodes MANN & WHITNEY et WALD & WOLFOWITZ.

MANN & WHITNEY s'appuie sur la somme des rangs (le rang des revenus) calculés sur chaque population (ceux qui ont un jardin et ceux qui n'en ont pas). La statistique n'est pas réellement interprétable, plus intéressant est la statistique Z qui suit asymptotiquement une loi normale centrée et réduite (Siegel & Castellan, pages 128 à 137). La statistique corrigée tient compte des ex-aequo en ajustant la variance estimée. Dans notre exemple, nous constatons qu'à un risque de 5% le revenu des ménages ne diffère pas selon qu'ils occupent ou pas une habitation avec jardin.

Mann-Whitney Comparison 1

Parameters

Sort results no

Results

		Value	Examples	Average	Rank sum	Statistics	Value	Z-Value	Proba
Income	Garden	No	220	3628.0909	32981.0	Mann-Whitney	8929.000000	0.194152	0.846057
		Yes	80	3647.7875	12169.0	MW corrected	8929.000000	0.194153	0.846056
		All	300	3633.3433	45150.0				

Computation time : 0 ms.
Created at 18/07/2005 17:09:36

WALD & WOLFOWITZ s'appuie sur le nombre de séquences pour détecter les différences entre les populations. Son utilisation est un peu dévoyée dans ce cas, cette statistique est surtout indiquée lorsque l'on veut détecter l'évolution aléatoire d'un indicateur (Siegel & Castellan, pages 58 à 64). Toutes observations sont mélangées et triées selon INCOME, puis étiquetées selon GARDEN, nous avons pu compter 111 séquences. La statistique Z suit asymptotiquement une loi normale centrée et réduite, à 5%, les données ne contredisent pas l'hypothèse nulle. Les résultats sont cohérents avec le test précédent.

Attention il s'agit le test de séquences est un test unilatéral à gauche, on constate une différence entre les deux populations si le nombre de séquences est significativement faible par rapport au nombre de séquences théoriques sous l'hypothèse nulle. Ce test est peu puissant (Siegel & Castellan, p.64).

Attribute_Y	Attribute_X	Description			Statistical test	
		Value	Examples	Average	Measure	Value
Income	Garden	No	220	3628.0909	Runs	111
		Yes	80	3647.7875	Z	-1.011359
		All	300	3633.3433	p-value	0.155922

Comparaison de K populations

Si l'on veut comparer maintenant K populations ($K > 2$), les tests ci-dessus ne sont pas applicables, nous devons nous tourner vers d'autres méthodes, en particulier le test de Kruskal & Wallis.

Plaçons un nouveau composant DEFINE STATUS à la racine du diagramme, mettez INCOME en TARGET, et HOUSE (3 modalités) en INPUT. Nous voulons vérifier que le revenu des ménages diffère ou pas selon le statut de leur habitation (« propriétaire de leur habitation », « location », « maison familiale »).

KRUSKAL & WALLIS est une généralisation du test de MANN & WHITNEY, les résultats peuvent être lus de la même manière, la statistique en revanche suit une loi du KHI-2. La statistique corrigée tient compte des ex-æquo (Siegel & Castellan, pages 206 à 212).

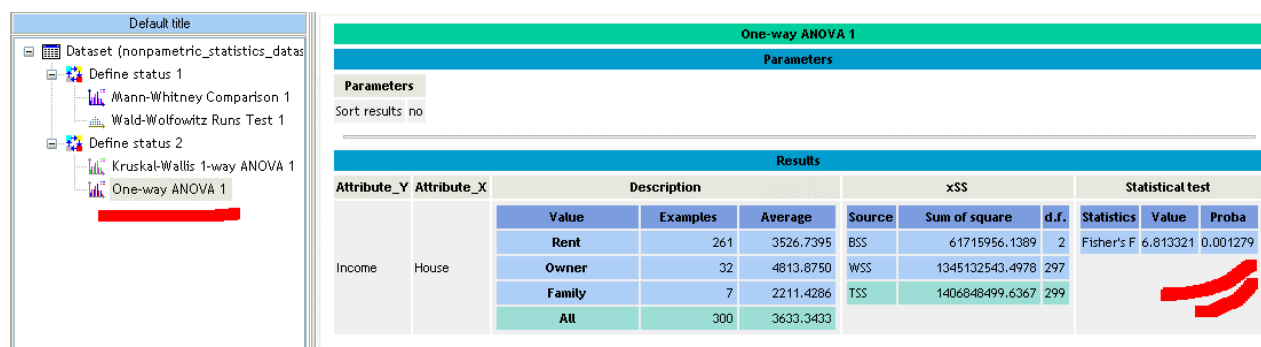
Attribute_Y	Attribute_X	Description				Statistical test		
		Value	Examples	Average	Rank sum	Statistics	Value	Proba
Income	House	Rent	261	3526.7395	38799.0	Kruskal-Wallis	9.318768	0.009472
		Owner	32	4813.8750	5814.0			
		Family	7	2211.4286	537.0	KW corrected	9.318880	0.009472
		All	300	3633.3433	45150.0			

Nous constatons qu'à 5%, le revenu des ménages diffère selon la situation par rapport à l'habitation. Dans le tableau de résultats, nous observons que ceux qui habitent dans un logement familial (FAMILY) ont un revenu moyen de #2211, ceux qui louent leur logement

(RENT) ont un revenu de #3526, et enfin les propriétaires de leur logement (OWNER) ont en moyenne un revenu de #4813.

Il existe bien une généralisation du test des séquences, c'est le test de MOOD (A.MOOD, « The distribution theory of runs », Ann. of Math. Stat., 11, pp. 367-392, 1940), mais cette méthode n'étant implémentée nulle part, il n'a pas été possible d'effectuer des comparaisons, nous l'avons introduit dans TANAGRA mais les résultats sont à prendre avec beaucoup de précautions.

En revanche, il est possible d'appliquer une technique paramétrique sur cette configuration : l'analyse de variance (ONE WAY ANOVA). Les deux tests produisent des résultats similaires.



One-way ANOVA 1										
Parameters										
Parameters										
Sort results: no										
Results										
Attribute_Y	Attribute_X	Description			xSS			Statistical test		
		Value	Examples	Average	Source	Sum of square	d.f.	Statistics	Value	Proba
Income	House	RENT	261	3526.7395	BSS	61715956.1389	2	Fisher's F	6.813321	0.001279
		OWNER	32	4813.8750	WSS	1345132543.4978	297			
		FAMILY	7	2211.4286	TSS	1406848499.6367	299			
		ALL	300	3633.3433						

Corrélation

Dans cette section, nous abordons une problématique différente : il s'agit de montrer que dans un couple, l'homme et la femme ont des salaires similaires ou, en d'autres termes, des personnes ayant des salaires comparables ont tendance à se mettre en ménage.

Une technique simple pour vérifier cette assertion est le calcul d'une corrélation entre le salaire des hommes et le salaire des femmes. Plaçons un composant DEFINE STATUS dans notre diagramme, mettons SALAIREHOMME en TARGET et SALAIREFEMME en INPUT. Deux composants qui peuvent servir dans ce cadre : le coefficient de corrélation de SPEARMAN et le tau de KENDALL.

Nous constatons qu'ils produisent des résultats très proches et nous emmènent à conclure, à 5%, qu'effectivement les salaires de l'homme et de la femme composant un couple sont significativement corrélés. Notons que, sur notre exemple, nous aurions obtenu la même conclusion avec le coefficient de corrélation linéaire (coefficient de Pearson), en revanche, les statistiques non-paramétriques détectent les relations non linéaires monotones, elles sont également robustes face aux données atypiques.

The screenshot displays three statistical result windows. Each window contains a table with the following data:

Y	X	r	t	Pr(> t)
MaleSalary	FemaleSalary	0.7383	18.8970	0.0000

Y	X	r	t	Pr(> t)
MaleSalary	FemaleSalary	0.5521	14.2609	0.0000

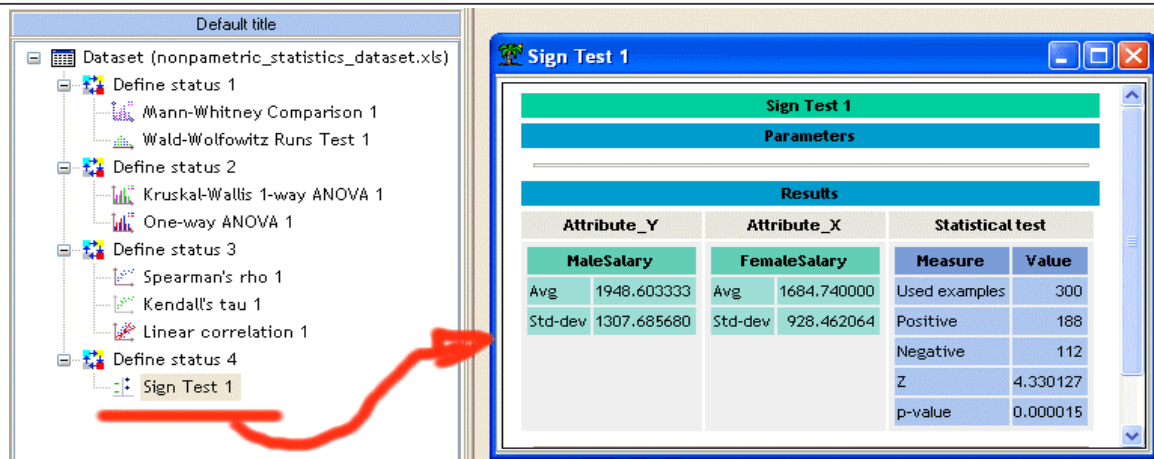
Y	X	r	t	Pr(> t)
MaleSalary	FemaleSalary	0.8784	31.7343	0.0000

Comparaisons sur échantillons appariés

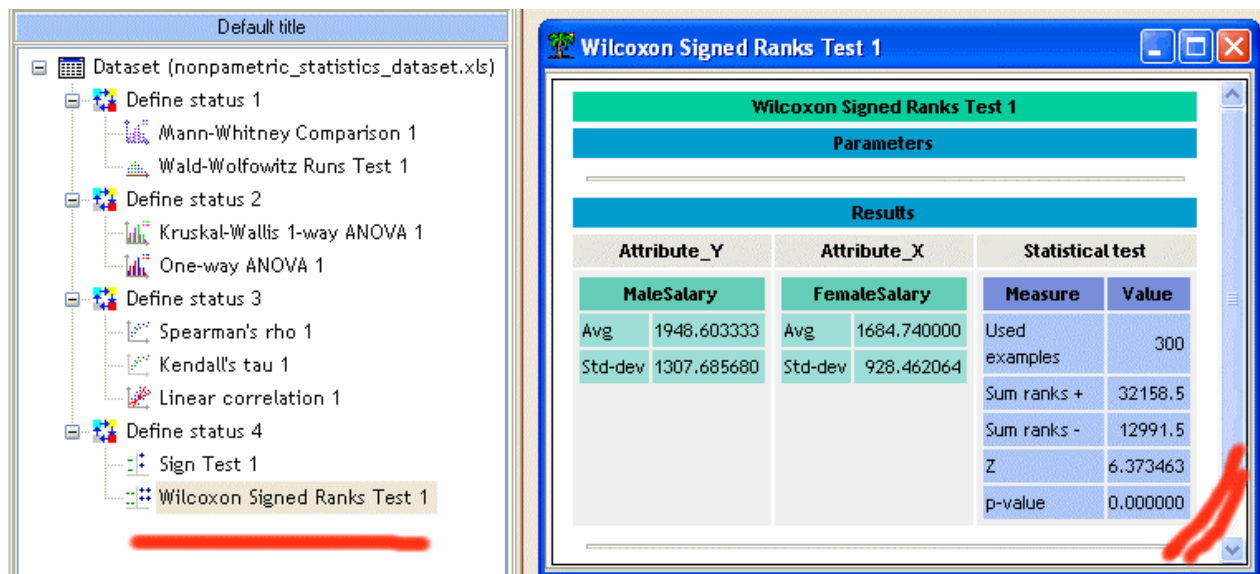
Dernier sujet que nous aborderons dans ce didacticiel, nous voulons vérifier que dans un ménage, l'homme a tendance à avoir un salaire plus élevé que celui de la femme.

Nous sommes typiquement dans une situation d'appariement, il ne s'agit surtout pas de calculer la moyenne des salaires des hommes, celle des femmes, et de les comparer. L'unité statistique est le ménage, et le concept qui nous intéresse est l'écart de salaire entre l'homme et la femme au sein du ménage. Plusieurs techniques non-paramétriques permettent de résoudre ce type de problème, il s'agit notamment du test des signes et du test de Wilcoxon pour échantillons appariés.

SIGN TEST consiste à compter le nombre de fois où le salaire des hommes est supérieur à celui des femmes, puis de confronter cette statistique avec la valeur théorique obtenue sous l'hypothèse nulle « égalité des salaires » (Siegel et Castellan, pages 80 à 87). Placez de nouveau le composant DEFINE STATUTS, mettez en TARGET la variable SALAIREHOMME, SALAIREFEMME en INPUT. Introduisez le composant SIGN TEST dans le diagramme. Les résultats montrent que sur les 300 ménages étudiés, l'homme a dans 188 cas, un salaire plus élevé que celui de la femme. L'indicateur Z suit une loi normale centrée réduite, le test est bilatéral. Dans notre cas, nous constatons qu'à 5%, le salaire de l'homme est en moyenne significativement supérieur à celui de la femme dans le ménage.



Le test des signes est très conservateur, en effet s'il tient compte du sens de l'écart, il ne tient pas compte de son amplitude. Le test de WILCOXON SIGNED RANK TEST est plus intéressant en ce sens car il utilise le rang des écarts (Siegel & Castellan, pages 87 à 95). Plaçons le dans notre diagramme, nous constatons que les résultats renforcent la conclusion ci-dessus.



Enfin, dernière alternative paramétrique à ces tests, le T de Student pour échantillons appariés utilise explicitement les écarts pour construire sa statistique. Il confirme également les résultats ci-dessus.

Paired T-Test 1

Parameters

Results

Attribute_Y		Attribute_X		Statistical test	
MaleSalary		FemaleSalary		Measure	Value
Avg	1948.603333	Avg	1684.740000	D avg.	263.863333
Std-dev	1307.685680	Std-dev	928.462064	D std-dev	662.561966
				T-test	6.897841
				p-value	0.000000

Computation time : 0 ms.
Created at 18/07/2005 19:33:28

Toutes les méthodes implémentées dans TANAGRA ont été validées de plusieurs manières : tout d'abord nous avons reproduit les exemples décrits dans notre ouvrage de référence, cela a permis de valider le détail des calculs ; par la suite, nous avons pris plusieurs fichiers benchmarks et nous avons comparé nos résultats avec les principaux logiciels commerciaux du marché.