1 Objectif

Tests non paramétriques de comparaison de (K > 2) populations. Modèle de localisation.

Les **tests de comparaison de populations** visent à déterminer si (K ≥ 2) échantillons proviennent de la même population au regard d'une variable d'intérêt (X). En d'autres termes, nous souhaitons vérifier que la distribution de la variable est la même dans chaque groupe. On utilise également l'appellation « tests d'homogénéité » dans la littérature.

Les tests **non paramétriques** lorsque l'on ne fait pas d'hypothèse sur la distribution de X, on parle aussi de tests « *distribution free* ».

Dans ce didacticiel, nous nous intéressons plus particulièrement à la configuration où la variable d'intérêt prend stochastiquement des valeurs plus élevées (ou plus faibles, ou simplement différentes) dans une des sous populations. On suppose que la différenciation se fait sur un décalage entre les caractéristiques de tendance centrale des distributions conditionnelles. On parle de modèle de localisation. Le **test de Kruskal-Wallis** est certainement celui qui vient immédiatement à l'esprit pour traiter ce type de problèmes. Nous verrons dans ce didacticiel que d'autres tests existent. Nous comparerons les résultats obtenus. Nous complèterons l'étude en procédant à des comparaisons multiples, on souhaite détecter les groupes qui diffèrent significativement les uns des autres.

Les aspects théoriques relatifs à ce didacticiel sont décrits dans un support de cours accessible en ligne http://eric.univ-lyon2.fr/~ricco/cours/cours/Comp_Pop_Tests_Nonparametriques.pdf.

2 Données

Les données proviennent du site de cours en ligne du Pr Richard Lowry du « Vassar College ». Nous traitons l'exemple utilisé pour illustrer le test de Kruskal-Wallis[1]. On a demandé à n = 21 personnes d'évaluer 3 types de vins (A, B et C) : n1 = 8 ont noté le premier type de vin 1, n2 = 7 pour le second et, n3 = 6 pour le troisième. On souhaite savoir si les notes attribuées sont significativement différentes d'un groupe à l'autre.

Il y a une grosse feinte dans l'expérimentation. En réalité, le vin est exactement le même quel que soit le groupe. C'est l'entretien d'évaluation, débouchant sur l'attribution de la note, qui a été mené de différentes manières. Il est enthousiaste pour le groupe A, un peu moins dans le groupe B, il est neutre dans le groupe C.

La variable d'intérêt est RATING. Elle va de 1 à 10, meilleure sera l'appréciation, plus élevée sera la note. Un aspect complémentaire intéressant de ce tutoriel serait d'étudier le comportement des méthodes paramétriques (ANOVA à 1 Facteur et WELCH ANOVA) sur ces mêmes données[2].

[1] http://faculty.vassar.edu/lowry/ch14a.html ; la démarche et les formules sont décrites en détail.

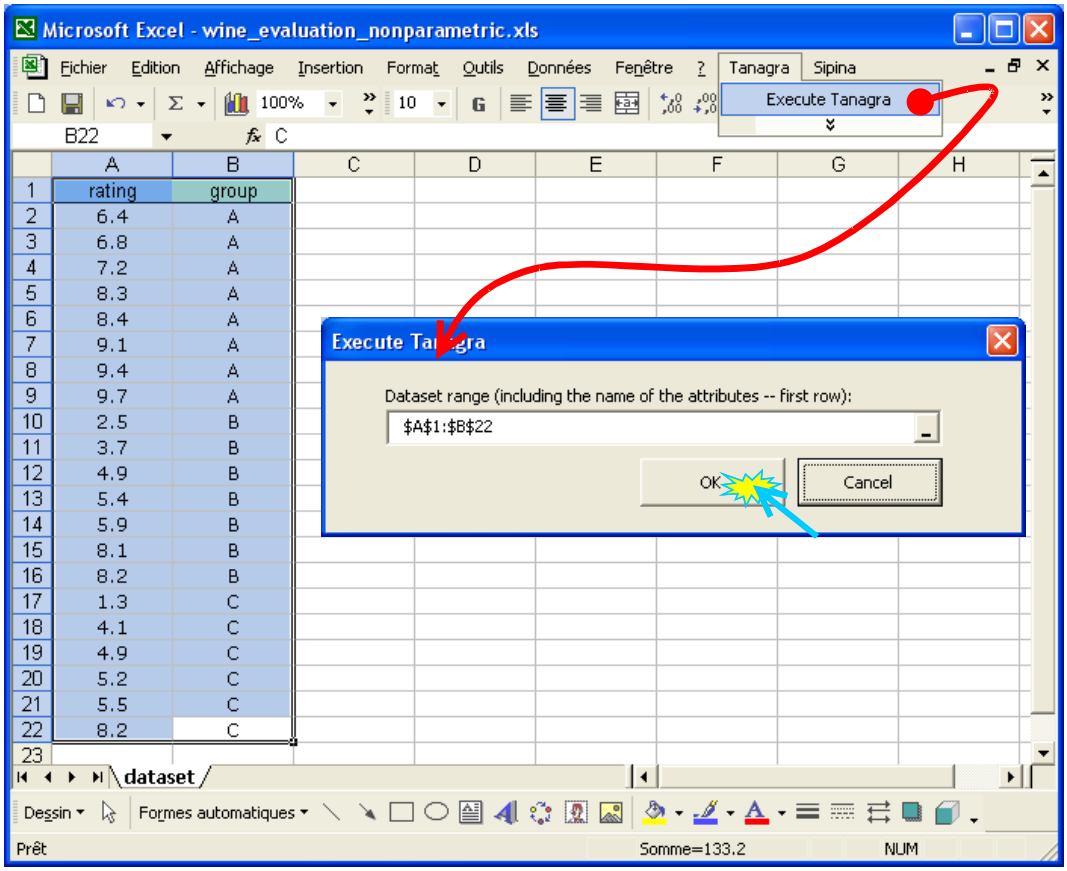
[2] Voir http://tutoriels-data-mining.blogspot.com/2008/07/comparaison-de-populations-tests.html pour la mise en œuvre de ces techniques dans TANAGRA.

3 Test de Kruskal-Wallis

Les données sont listées dans le fichier **wine_evaluation_nonparametric.xls**[3]. Les observations sont décrites par 2 variables, la note attribuée (RATING) et le groupe d'appartenance du goûteur (GROUP : « A », « B » ou « C »).

3.1 Importation des données

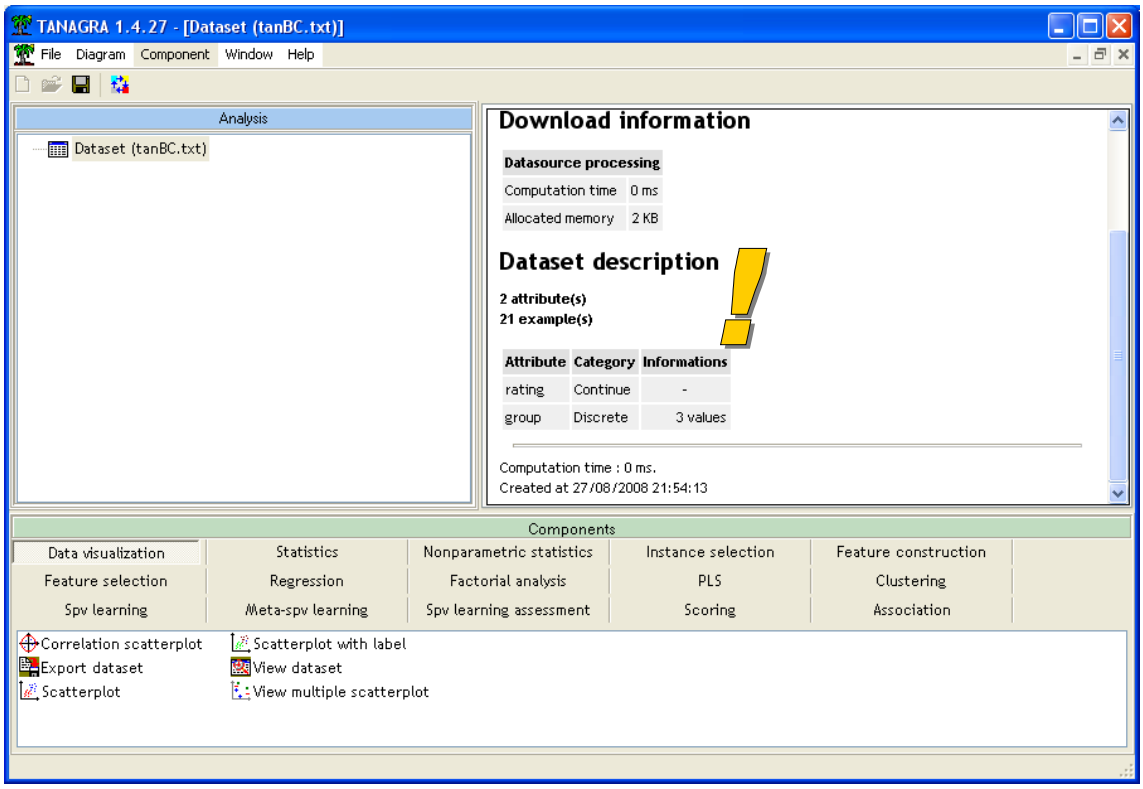
Le plus simple pour lancer Tanagra et charger les données est d'ouvrir le fichier XLS dans le tableur EXCEL. Nous sélectionnons la plage de données. La première ligne doit correspondre au nom des variables. Puis nous activons le menu TANAGRA / EXECUTE TANAGRA qui a été installé avec la macro complémentaire TANAGRA.XLA[4]. Une boîte de dialogue apparaît. Nous vérifions la sélection. Si tout est en règle, nous validons en cliquant sur le bouton OK.



TANAGRA est automatiquement lancé. Un nouveau diagramme est créé. Nous devons disposer de 21observations et 2 variables.

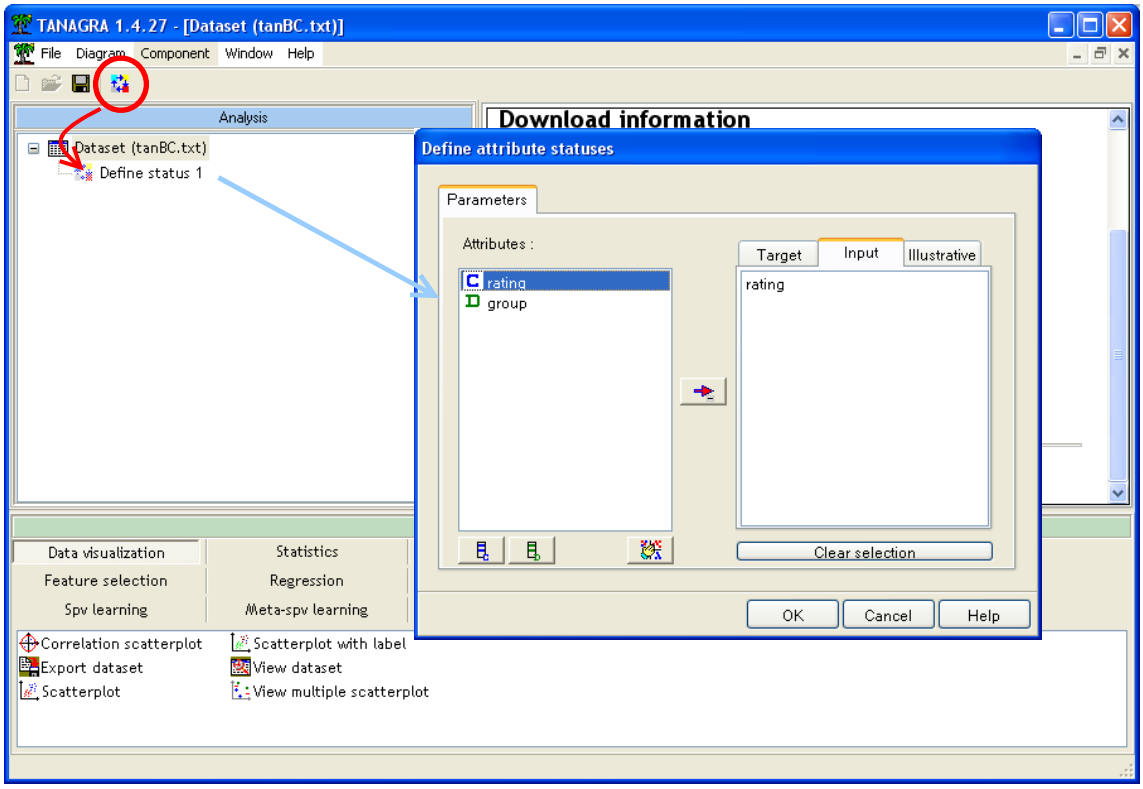
[3] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/wine_evaluation_nonparametric.xls

[4] Voir http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-macro.html concernant l'installation et l'utilisation de la macro complémentaire TANAGRA.XLA.

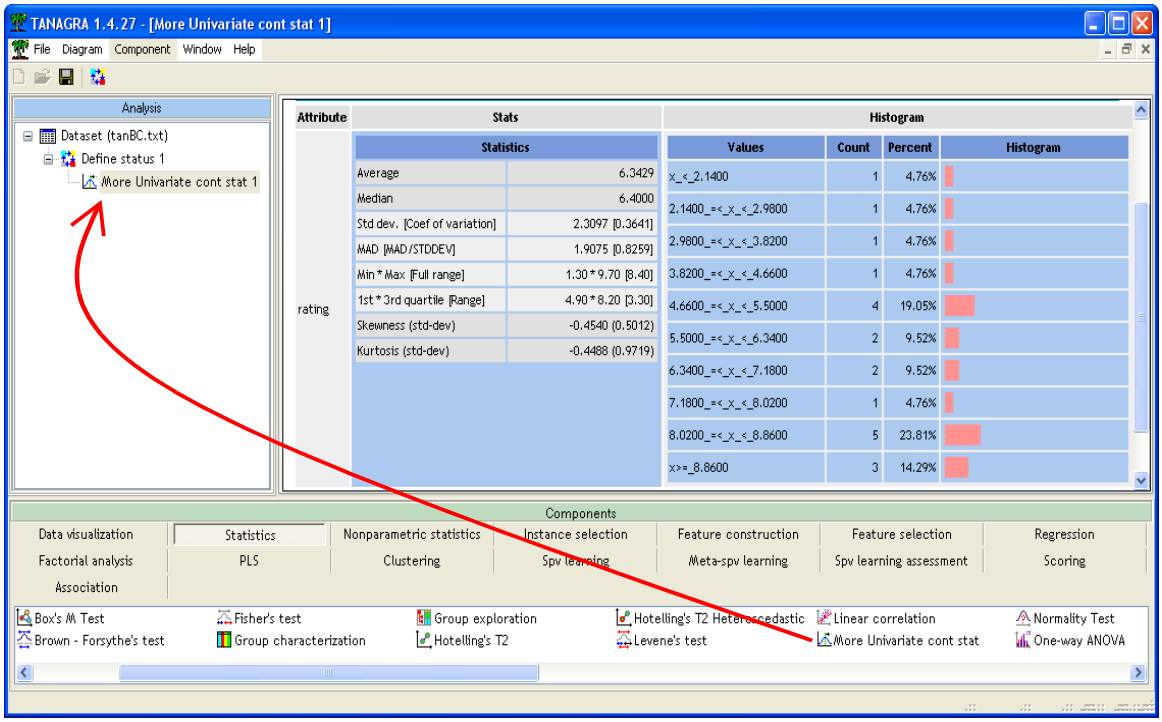


3.2 Statistiques descriptives

Premier préalable, toujours indispensable, nous calculons quelques indicateurs de statistique descriptive sur les données, ne serait-ce que pour en vérifier l'intégrité. Pour ce faire, nous insérons le composant DEFINE STATUS via le raccourci dans la barre d'outils dans le diagramme, nous plaçons en INPUT la variable RATING

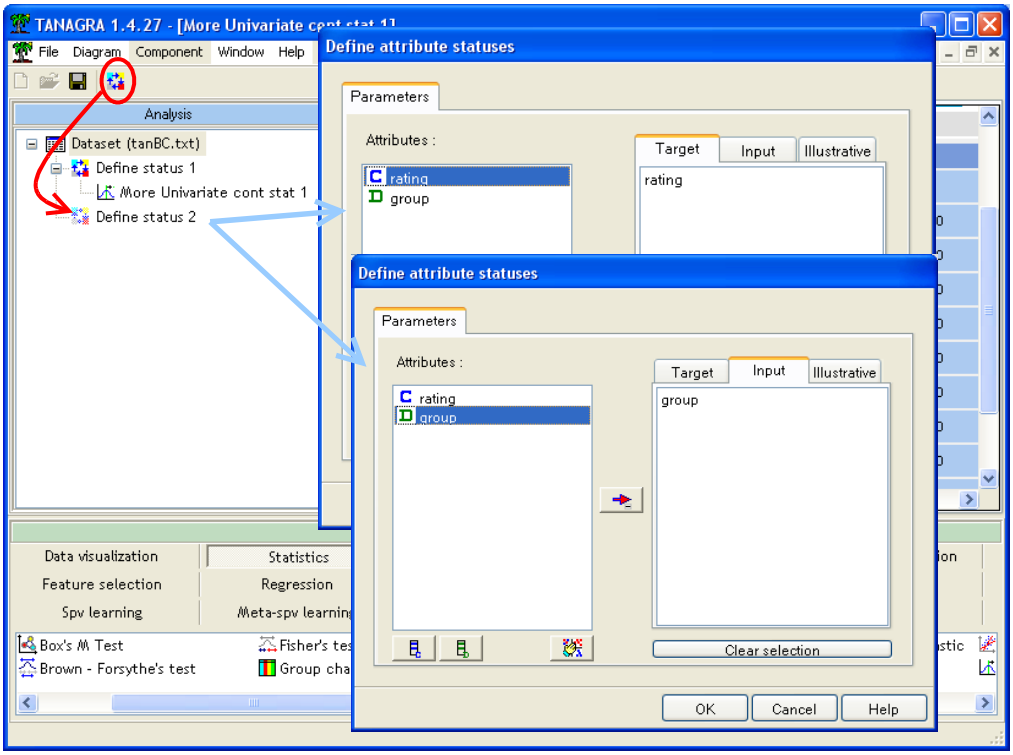


Puis nous insérons le composant MORE UNIVARIATE STAT (onglet STATISTICS). Nous obtenons le résultat suivant en cliquant sur le menu VIEW. Nous nous bornerons à observer que les notes varient de min = 1.30 à max = 9.70.

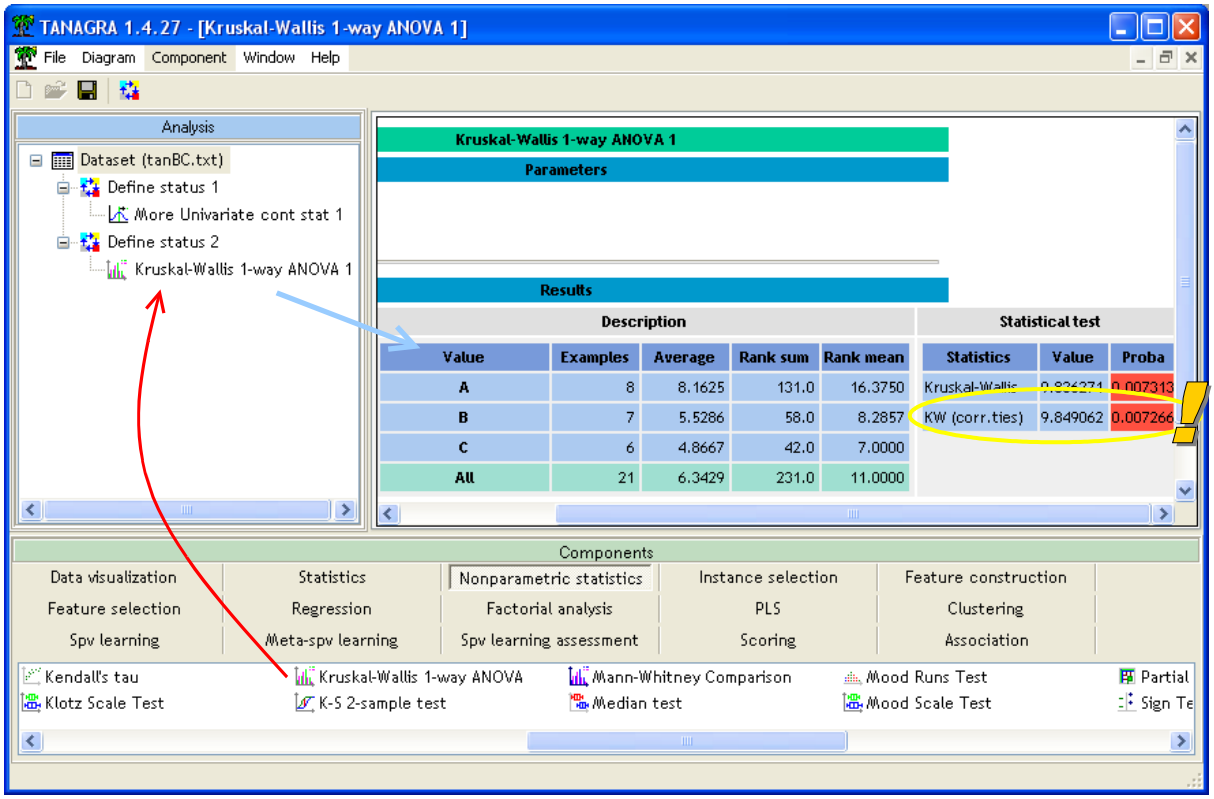


3.3 Test de Kruskal-Wallis

Nous voulons tester la significativité des écarts à l'aide du test de Kruskal-Wallis. Nous insérons de nouveau le composant DEFINE STATUS dans le diagramme. Nous plaçons maintenant en TARGET la variable RATING, en INPUT la variable GROUP.



Nous désirons savoir si les écarts de notations sont stochastiquement significatifs au risque 5% Nous insérons alors le composant KRUSKAL-WALLIS 1-WAY ANOVA (onglet NONPARAMETRIC STATISTICS). Nous cliquons sur le menu contextuel VIEW pour obtenir les résultats.



TANAGRA affiche la statistique de test, sans et avec prise en compte des ex-aequo (TIES). C'est cette dernière qui nous intéresse en premier lieu, nous avons KW = 9.849062. Sous l'hypothèse nulle d'égalité des distributions conditionnelles, elle suit une loi du χ^2 à (K - 1 = 3 – 1 = 2) degrés de liberté. La probabilité critique du test est p = 0.007266. Elle est plus petite que le risque nominal que l'on s'est fixé (5%). On conclut que le niveau de notation dans au moins un des groupes est différent des autres (ou d'un autre).

3.4 Détecter la source des écarts

Lorsque nous aboutissons au rejet de l'hypothèse nulle, l'étape suivante est souvent la détection des écarts significatifs. On procède à des comparaisons multiples en opposant 2 à 2 les sous-groupes. On considère qu'un écart entre les rangs moyens du groupe i et j est significatif si nous observons la situation suivante[5]

$$\left|\bar{r}_i - \bar{r}_j\right| \geq u_{1-\alpha}\sqrt{\frac{n(n+1)}{12}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

où

- n est l'effectif total, n_i (resp. n_j) est l'effectif du groupe i (resp. j) ;

[5] S. Siegel, J. Castellan, « Nonparametric Statistics for the Behavioral Sciences », McGraw-Hill, Inc., 1988 ; pp.213-214.

- $\bar{r}_i$ (resp. $\bar{r}_i$) est le rang moyen du groupe i (resp. j) ;
- α est le risque global choisi pour le test de Kruskal-Wallis (5% dans notre exemple) ;
- $a = \dfrac{\alpha}{K(K-1)}$ est le risque pour les tests individuels ;
- u_{1-a} est le quantile d'ordre (1-a) de la loi normale centrée réduite.

Détaillons les calculs pour l'opposition entre le groupe A et B :

- $n = 21$, $n_A = 8$ et $n_B = 7$;
- $\bar{r}_A = 16.375$, $\bar{r}_B = 8.2857$, $|\bar{r}_A - \bar{r}_B| = |16.375 - 8.2857| = 8.089$;
- $\alpha = 0.05 \rightarrow a = \dfrac{0.05}{3(3-1)} = 0.008 \rightarrow u_{1-0.0083} = 2.3940$
- $u_{1-a}\sqrt{\dfrac{n(n+1)}{12}\left(\dfrac{1}{n_i}+\dfrac{1}{n_j}\right)} = 2.3940\sqrt{\dfrac{21(21+1)}{12}\left(\dfrac{1}{8}+\dfrac{1}{7}\right)} = 7.688$

Puisque $|\bar{r}_i - \bar{r}_j| = 8.089 > 7.688$, nous pouvons considérer que les notes attribuées par les goûteurs sont différents dans les groupes A et B.

Nous avons réalisé les calculs pour les autres oppositions. Il apparaît que les écarts sont significatifs pour les oppositions A vs. B, A vs. C.

Description

Value	Examples	Average	Rank sum	Rank mean
A	8	8.1625	131	16.375
B	7	5.5286	58	8.2857
C	6	4.8667	42	7
All	21	6.3429	231	11

Statistical test

Statistics	Value	Proba
Kruskal-Wallis	9.836271	0.00731
KW (corr.ties)	9.849062	0.00727

n	21
K	3

alpha	0.05
a	0.0083

u	2.3940

Group.1	Group.2	|Difference|	Crit.Level	Significant
A	B	8.089	7.688	yes
A	C	9.375	8.022	yes
B	C	1.286	8.264	no

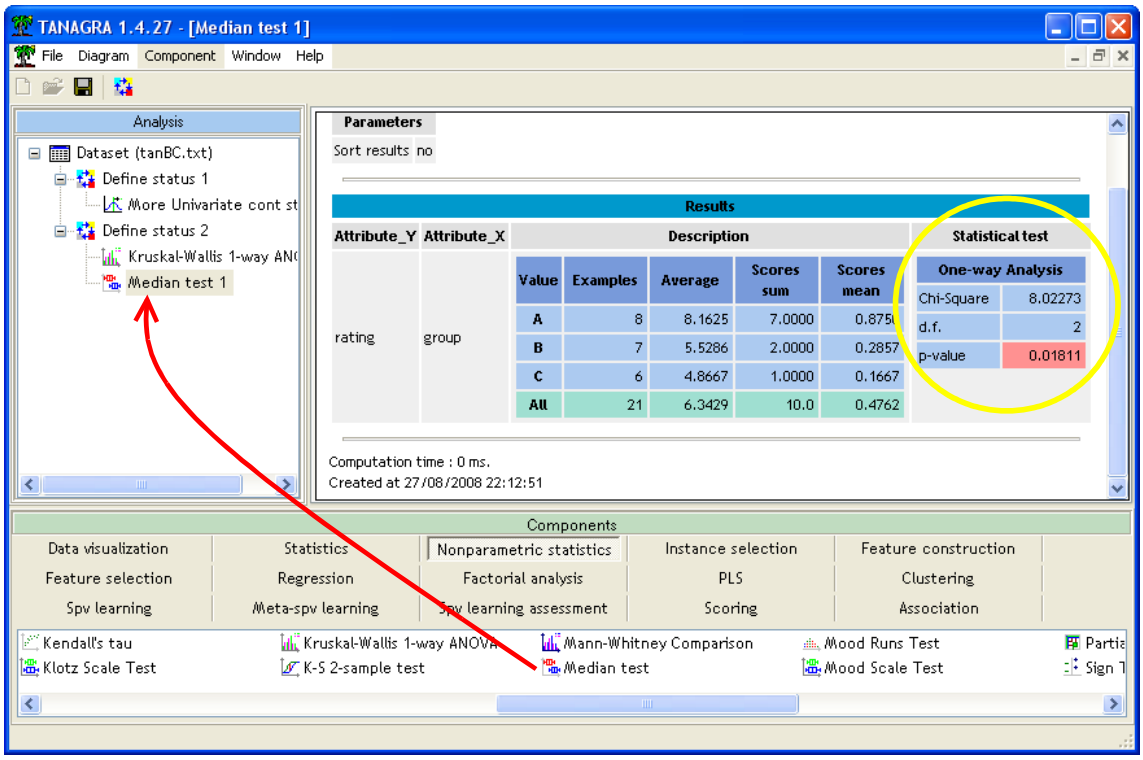
4 D'autres tests

Le test de Kruskal-Wallis est la procédure non paramétrique la plus populaire de comparaison de K échantillons indépendants. Mais ce n'est pas la seule. D'autres techniques existent, peu connues et/ou peu programmées dans les logiciels. Nous en présentons quelques unes dans cette section.

4.1 Le test des médianes

Ce test compare explicitement les médianes conditionnelles. Il est généralement moins puissant que le test de Kruskal-Wallis c.-à-d. conclut un peu trop souvent à la compatibilité des données avec l'hypothèse nulle. Il y a cependant des configurations où il doit être préféré à ce dernier, notamment lorsque les queues de distributions sont importantes[6].

Nous insérons le composant MEDIAN TEST (onglet NONPARAMETRIC STATISTICS) dans le diagramme. Nous cliquons sur le menu VIEW.



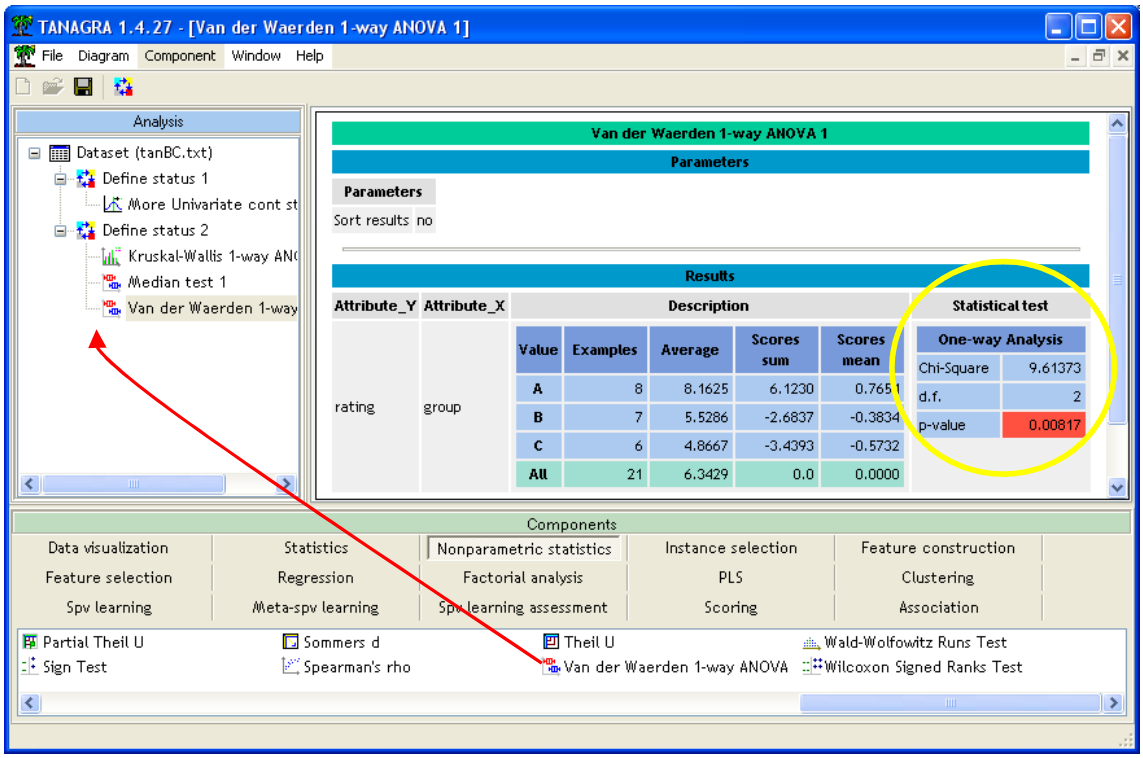
La statistique du test est $\chi^2 = 8.02273$. Avec un degré de liberté égal à 2, la probabilité critique est $p = 0.01811$. Au risque 5%, nous pouvons rejeter l'hypothèse nulle. La décision est toutefois moins tranchée qu'avec le test de Kruskal-Wallis.

4.2 Le test de Van der Waerden

Le test de Van der Waerden transforme les rangs en quantile de la loi normale, que l'on nomme scores normaux ou codes normaux. Cette approche est particulièrement indiquée lorsque la distribution sous-jacente des données est proche de la loi normale.

Nous insérons le composant VAN DER WAERDEN 1-WAY ANOVA dans le diagramme.

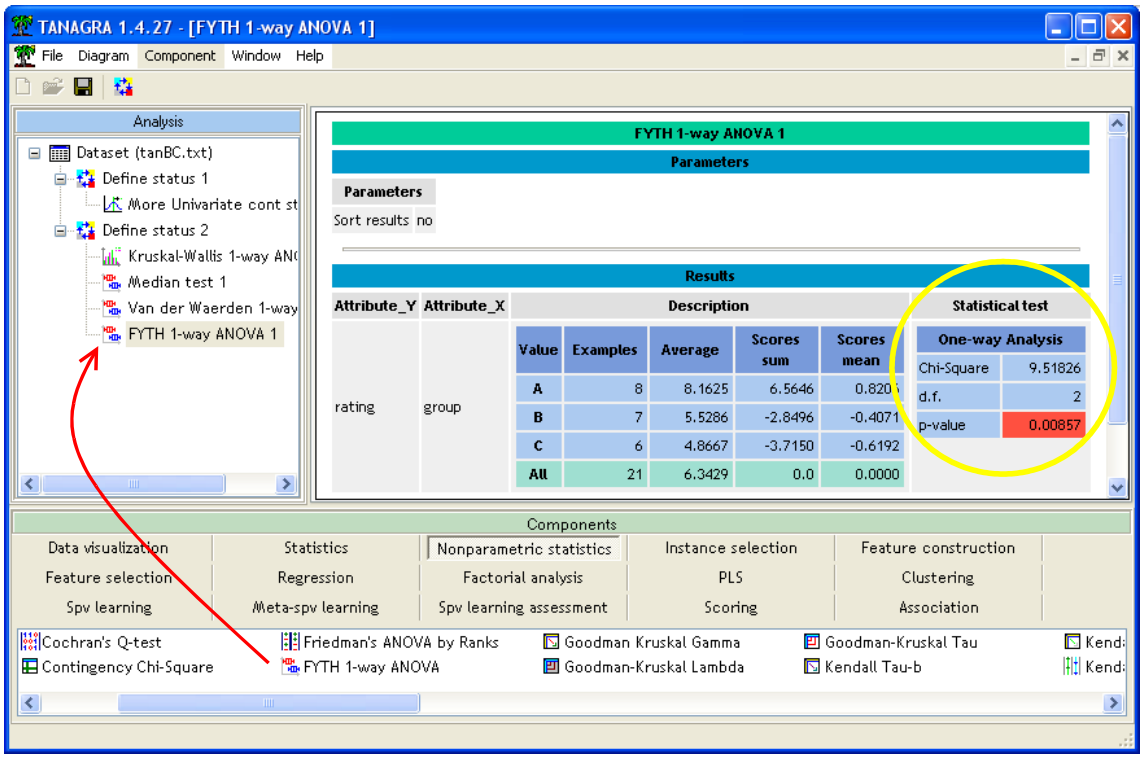
[6] Voir http://v8doc.sas.com/sashtml/stat/chap47/sect17.htm



La statistique du test est $\chi^2 = 9.61373$. Avec un degré de liberté égal à 2, la probabilité critique est $p = 0.00817$, très proche de celui proposé par le test de Kruskal-Wallis. Ce n'est pas étonnant compte tenu des caractéristiques de ces méthodes.

4.3 Le test de Fisher-Yates-Terry-Hoeffding

Le test de Fisher-Yates-Terry-Hoeffding est une variante très proche de Van der Waerden, basée sur des codes normaux. Nous insérons le composant FYTH 1-WAY ANOVA dans le diagramme.



La statistique du test est χ^2 = 9.51826. Avec un degré de liberté égal à 2, la probabilité critique est p = 0.00857.

5 Conclusion

Les composants MEDIAN TEST, VAN DER WAERDEN 1-WAY ANOVA et FYTH 1-WAY ANOVA sont opérationnels pour la comparaison de K = 2 échantillons indépendants. Un affichage supplémentaire est proposé, il s'agit de la statistique de test standardisée, asymptotiquement normale.

Le test de Kruskal-Wallis ne peut pas fonctionner pour la comparaison de K = 2 populations. On lui substituera le test de Mann et Whitney c.-à-d. le composant MANN-WHITNEY COMPARISON.