1 Objectif

Tests non paramétriques de comparaison de 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* ».

De manière générique, le test de **Kolmogorov-Smirnov** consiste à comparer les fonctions de répartition empiriques (CDF : *cumulative distribution function*, en anglais). Dans ce cas, on cherche toute forme de différenciation entre les distributions.

On peut approfondir l'analyse en qualifiant la forme de la différenciation. Une approche très usitée consiste à déterminer si les valeurs de la variable d'intérêt sont stochastiquement plus élevés (plus faibles, ou tout simplement différents) dans un des sous échantillons. Le test de **Wilcoxon-Mann-Whitney** est certainement la technique la plus populaire, nous verrons dans ce didacticiel que d'autres tests non paramétriques peuvent être utilisés.

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 de l'Université Penn State de Pennsylvanie « STAT 500 – Applied Statistics »[1]. Nous nous intéressons à la leçon n°10 qui traite de la comparaison de moyennes. Il s'agit d'évaluer les performances de 2 machines, une ancienne et une nouvelle, lors de l'empaquetage de cartons. La variable d'intérêt est la durée de l'opération.

Les données semblent compatibles avec une distribution normale, les tests paramétriques sont à privilégier dans ce cas. Le site d'ailleurs détaille les résultats du test de Student de comparaison de moyenne. La statistique du test est t = -3.40, l'écart est très significatif avec une probabilité critique (p-value) p = 0.0032 pour un test bilatéral.

Un aspect intéressant de ce tutoriel sera d'étudier le comportement les tests non paramétriques sur ces données, et de confronter les résultats avec celui du test de Student.

3 Importation des données et création d'un diagramme

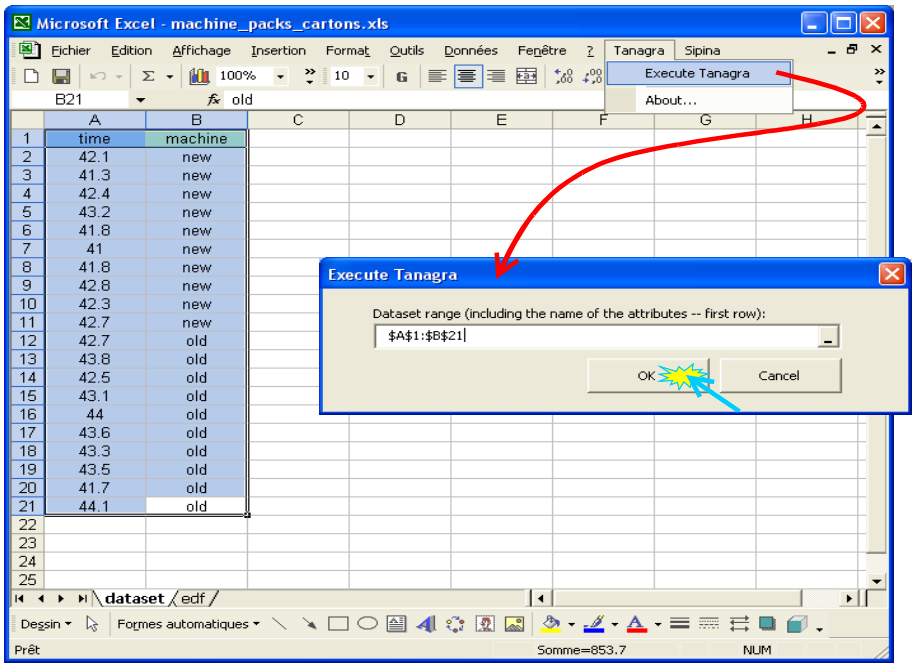
Les données sont listées dans le fichier **machine_packs_cartons.xls**[2]. Les observations sont décrites par 2 variables, le temps de traitement (TIME, en secondes) et le type de machine (MACHINE ; « new » ou « old »).

[1] http://www.stat.psu.edu/online/development/stat500_spss/index.html ; Lesson 10.

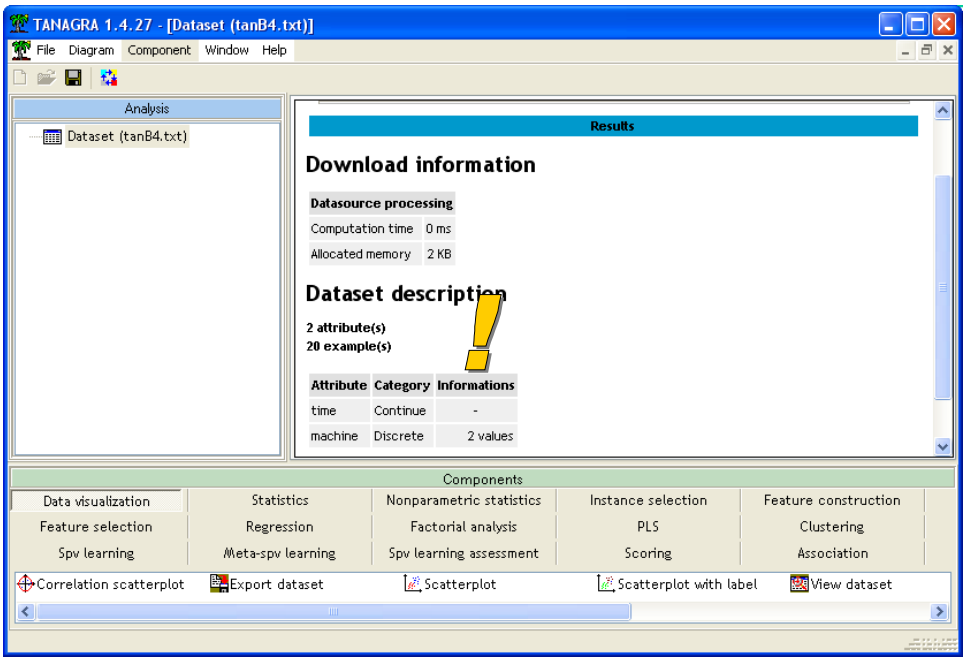
[2] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/machine_packs_cartons.xls

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[3]. Une boîte de dialogue apparaît. Nous vérifions la sélection. Si tout est en règle, nous validons en cliquant sur le bouton OK.



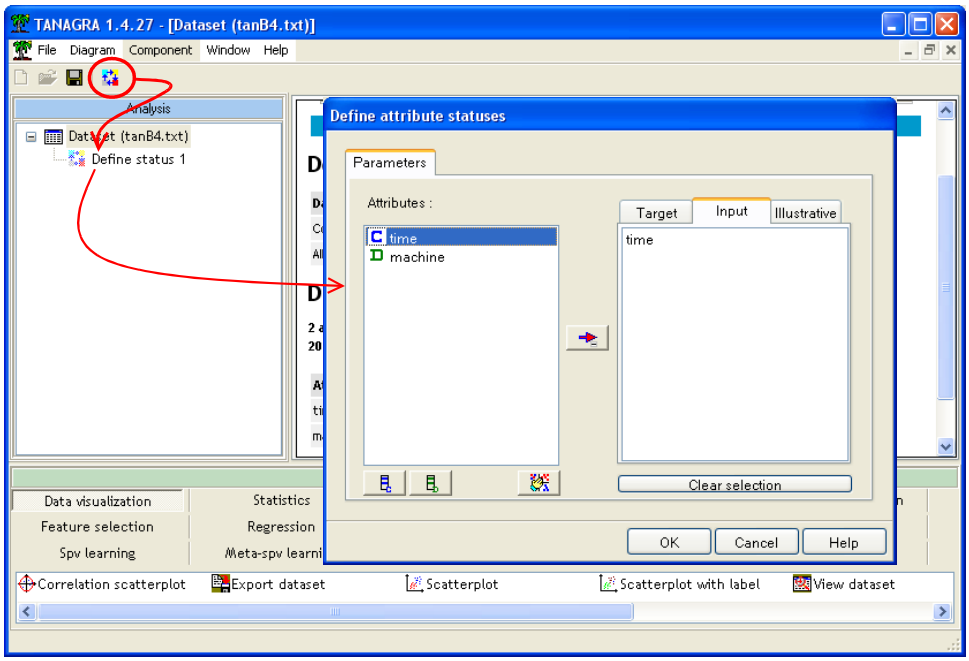
TANAGRA est automatiquement lancé. Un nouveau diagramme est créé. Nous devons disposer de 20 observations et 2 variables.



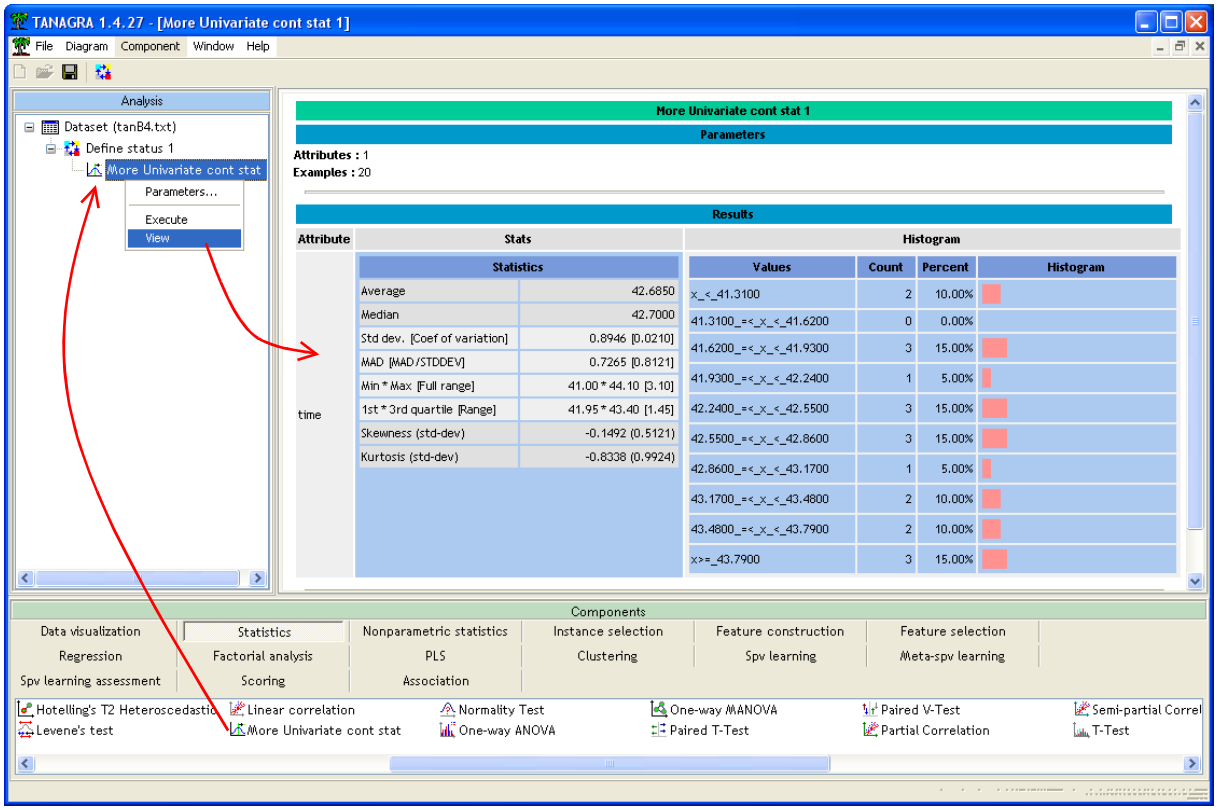
[3] Voir http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-macro.html concernant l'installation et l'utilisation de la macro complémentaire TANAGRA.XLA.

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 TIME.



Puis nous insérons le composant MORE UNIVARIATE STAT (onglet STATISTICS). Nous obtenons le résultat suivant en cliquant sur le menu VIEW.



Il n'y a pas de commentaires particuliers à émettre à ce stade. Tout juste remarquerons nous qu'il semble y avoir 2 pics (modes) dans la distribution : l'une vers 42.5, l'autre autour de 43.5. Reste à savoir si nous pouvons les associer aux groupes « new » et « old » relatifs au type de machine.

4 Comparaison des fonctions de répartition

Ce type de test confronte les fonctions de répartition empiriques conditionnelles. Nous produisons graphiquement (Figure 1). Apparemment, il semble y avoir un décalage fort entre les distributions. Si l'on s'intéresse aux médianes par exemple, pour F(X) = 0.5, nous constatons que celle des nouvelles machines est égale à 42.1, celle des anciennes, 43.3.

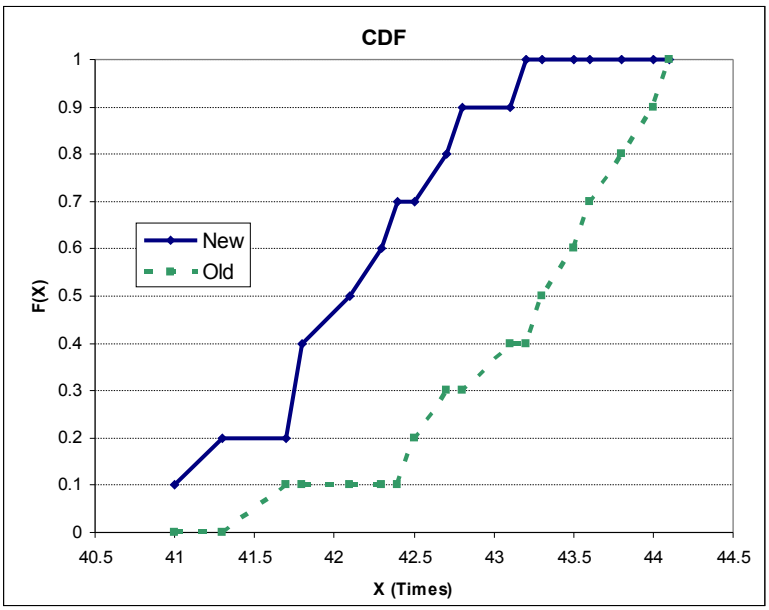
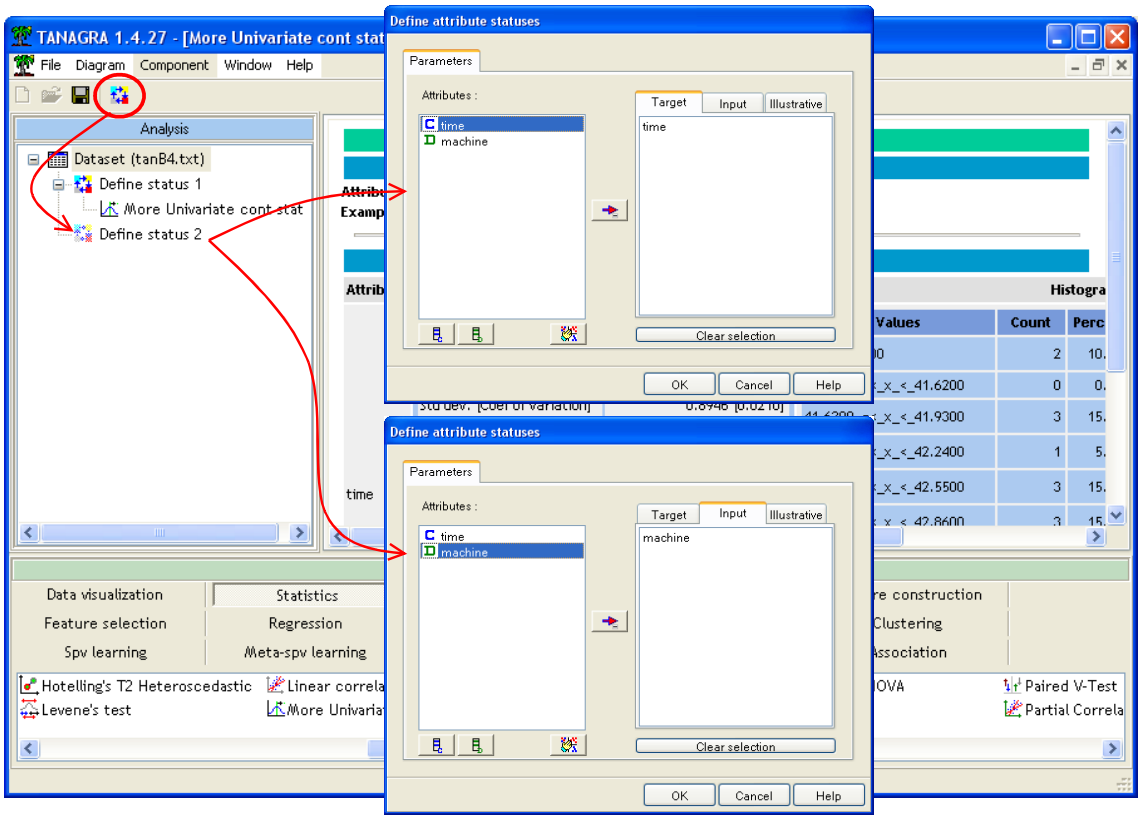
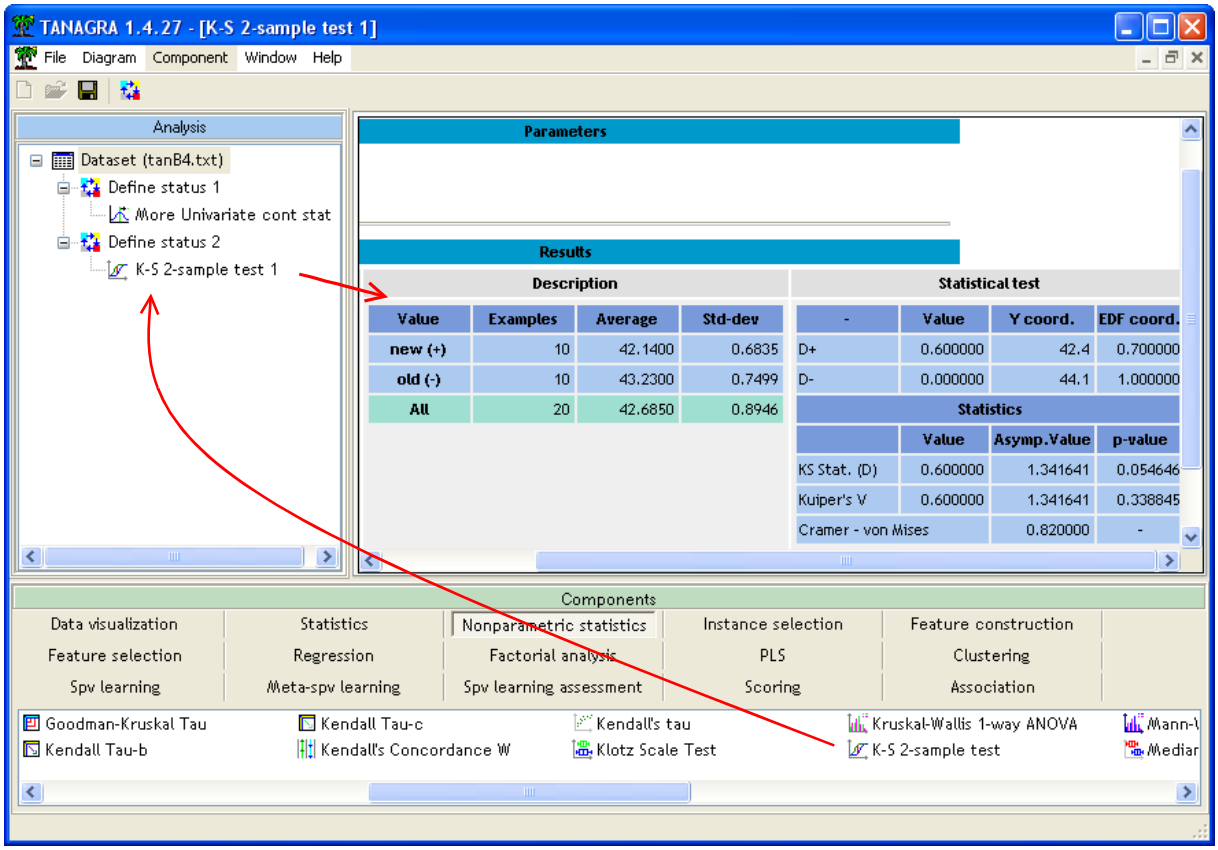


Figure 1 - Fonctions de répartition empiriques conditionnelles

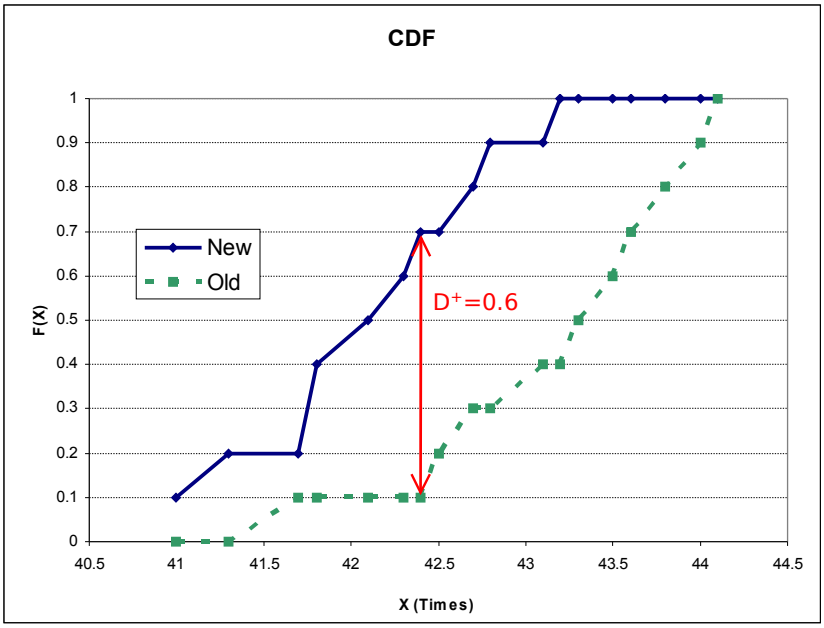
Essayons de confirmer cela statistiquement en utilisant le test de Kolmogorov-Smirnov. Nous insérons le composant DEFINE STATUS. Nous plaçons en TARGET la variable d'intérêt TIME ; en INPUT, MACHINE, la variable définissant les sous populations.



Nous introduisons ensuite le composant K-S 2-SAMPLE TEST (onglet NONPARAMETRIC STATISTICS). Nous cliquons sur le menu contextuel VIEW pour obtenir les résultats.



TANAGRA a défini comme groupe de référence (positive) la modalité « MACHINE = NEW ». La différence entre les fonctions de répartitions n'est jamais négative (D- = 0) c.-à-d. la fonction de répartition de OLD n'est jamais au dessus de celle de NEW. L'écart positif est à son maximum (D+ = 0.6) lorsque X = 42.4, avec $F_+(X) = 0.7$.



La statistique de Kolmogorov-Smirnov est D = max(D+ ; D-) = 0.6. La valeur utilisée pour calculer la probabilité critique à l'aide de la distribution asymptotique est $d = \sqrt{\frac{10 \times 10}{10 + 10}} \times D = 1.341641$. La probabilité critique est p = 0.054646.

Au niveau de risque 5%, nous dirons que les données sont compatibles avec l'hypothèse nulle d'égalité des fonctions de répartition.

Nous devons relativiser ce résultat néanmoins. Tout d'abord, nous sommes à la lisière de la région critique, une ou deux observations supplémentaires pourraient faire basculer la décision. Ensuite, l'écart visuel des fonctions de répartition conditionnelles dans la représentation graphique (Figure 1) laisse à penser qu'il y a quand même une différenciation, même si le test semble dire le contraire. Enfin, nous savons que le test de Kolmogorov-Smirnov est un test omnibus qui cherche toute forme de différenciation, il est de fait très peu puissant avec un risque de seconde espèce élevé (accepter l'hypothèse nulle alors que l'hypothèse alternative est vraie).

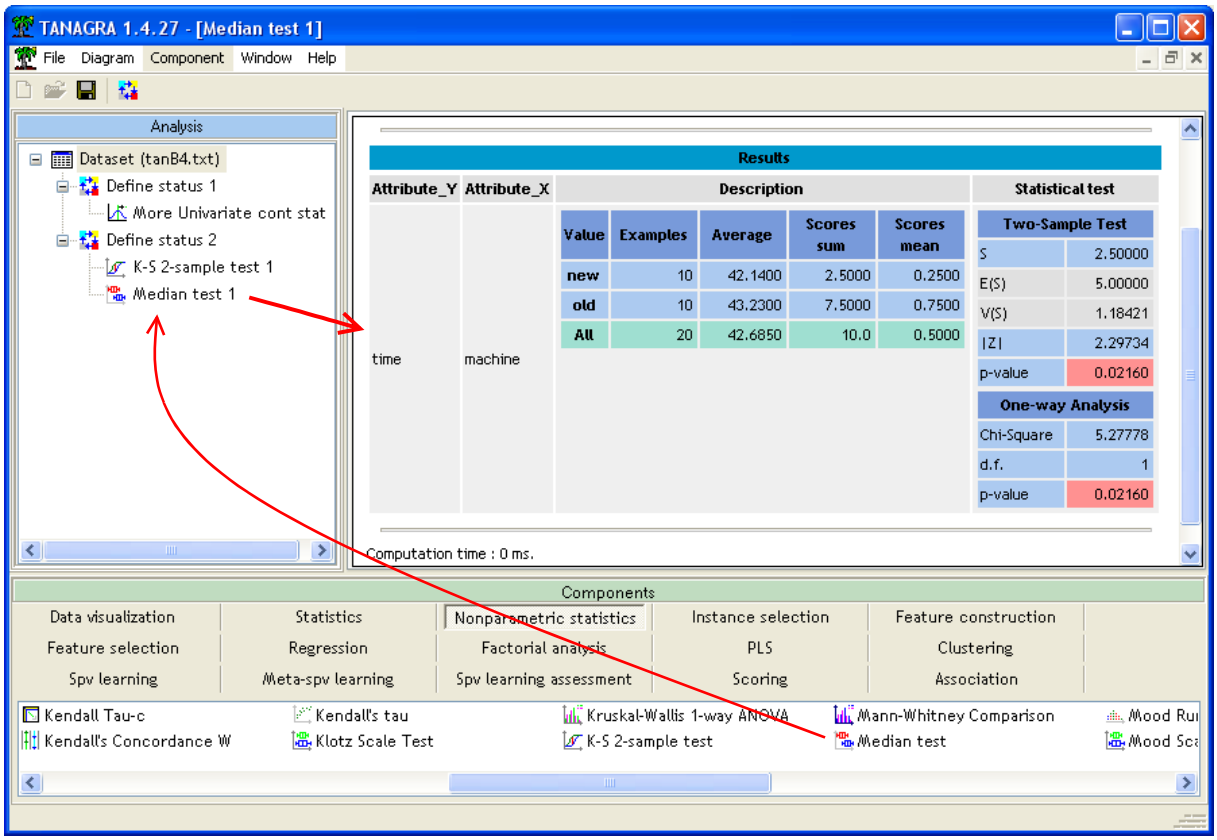
5 Comparaison selon un modèle de localisation

Il nous faut donc utiliser un test plus spécifique, qui permet de caractériser le décalage entre les caractéristiques de tendance centrale des distributions. On parle de modèle de localisation.

5.1 Test de la médiane

Ce test compare les médianes des deux distributions conditionnelles. Nous avions suggéré cette idée plus haut en mettant en parallèle les médianes observées. Il est temps maintenant de passer par une procédure rigoureuse.

Nous insérons le composant MEDIAN TEST (NONPARAMETRIC STATISTICS) dans le diagramme. Nous obtenons les résultats suivants.



La statistique de test est S = S1 = 2.5. Elle correspond à la somme des scores associés au premier sous échantillon (MACHINE = NEW). Pour le second groupe, nous avons S2 = 7.5. Pour évaluer la significativité de la différence, TANAGRA produit l'espérance et la variance de S sous l'hypothèse nulle, avec E(S) = 5.0 et V(S) = 1.18421. La statistique standardisée Z est alors égale à

$$|Z| = \frac{|S - E(S)|}{\sqrt{V(S)}} = \frac{|2.5 - 5.0|}{\sqrt{1.18421}} = 2.29734$$

Elle suit asymptotiquement une loi normale. La probabilité critique du test est p = 0.0216. A l'évidence, au niveau de risque 5%, les médianes du temps de traitement des nouvelles et anciennes machines sont significativement différentes.

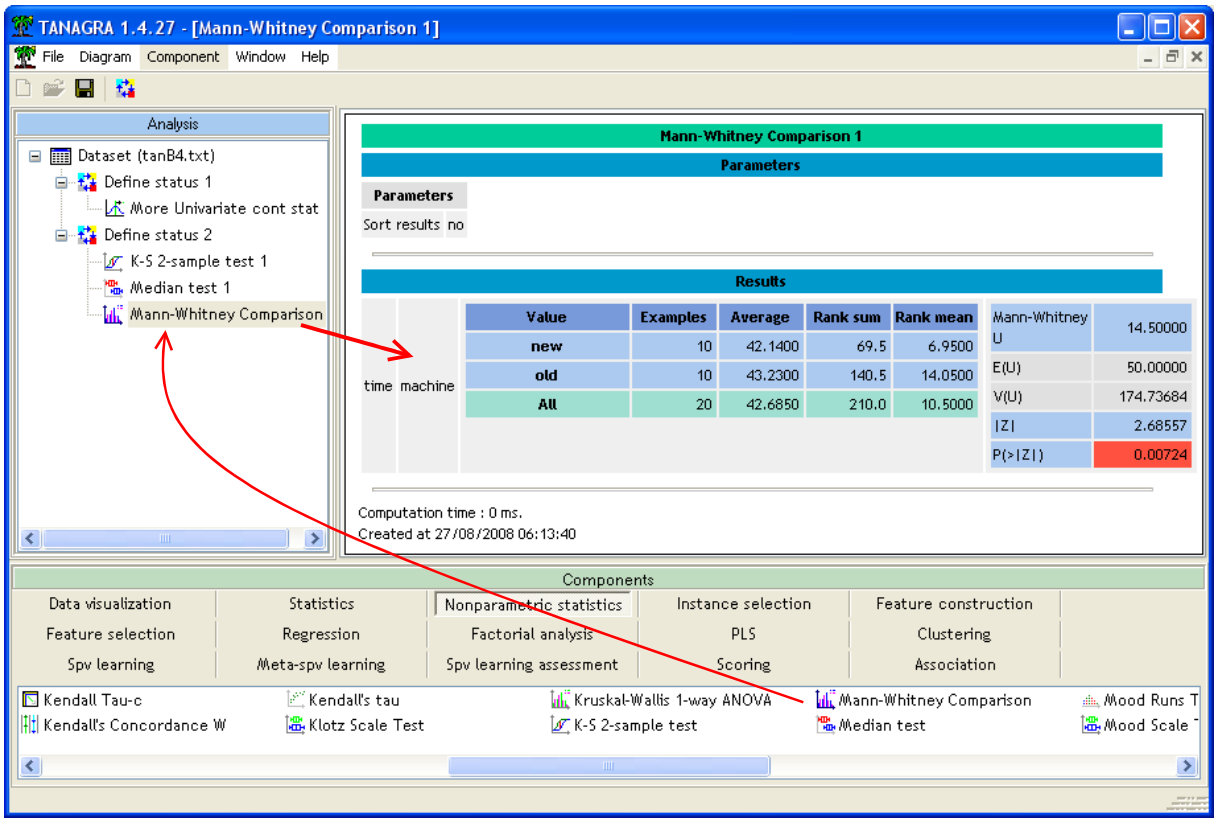
Le test peut être généralisée à la comparaison de (K ≥ 2) médianes. TANAGRA produit automatiquement la statistique du χ^2 à (K-1) degrés de liberté. Dans notre exemple, avec le cas particulier K=2, nous avons la relation $Z^2 = (2.29734)^2 = 5.27778 = \chi^2$. Les « p-value » sont strictement identiques.

5.2 Test de Wilcoxon-Mann-Whitney

Le test de la médiane est certainement plus adapté que le test de Kolmogorov-Smirnov pour notre problème. Il caractérise mieux le décalage entre les distributions. Pourtant, il n'est pas le plus puissant dans notre contexte. En effet, il cherche uniquement à positionner les individus par rapport à la médiane, il n'utilise le positionnement relatif des observations.

Le test non paramétrique le plus indiqué si nous souhaitons caractériser une différenciation selon la caractéristique de tendance centrale des distributions est le test de Wilcoxon-Mann-Whitney basé

sur les rangs. Nous insérons le composant MANN-WHITNEY COMPARISON (onglet NONPARAMETRIC STATISTICS) dans notre diagramme.



La somme des rangs pour le premier groupe MACHINE = NEW, qui sert de référence, est S1 = 69.5 ; pour le second, S2 = 140.5. La statistique du test est définie par

$$U = S_1 - \frac{n_1(n_1+1)}{2} = 69.5 - \frac{10(10+1)}{2} = 14.5$$

TANAGRA fournit E(U) = 50.0 et V(U) = 174.73684 sous l'hypothèse nulle. La statistique standardisée est

$$|Z| = \frac{|U - E(U)|}{\sqrt{V(U)}} = \frac{|14.5 - 50.0|}{\sqrt{174.73684}} = 2.68557$$

La probabilité critique est p = 0.00724.

L'écart qui semblait évident visuellement (Figure 1) est clairement confirmé par le test statistique cette fois-ci. Nous nous rapprochons des résultats du test de Student. Ce dernier paraissait le plus approprié compte tenu des vérifications préalables mises en place sur notre site de référence[4] (distributions conditionnelles gaussiennes, égalité des variances dans les sous-groupes). La conclusion du test de Wilcoxon-Mann-Whitney est en parfait accord.

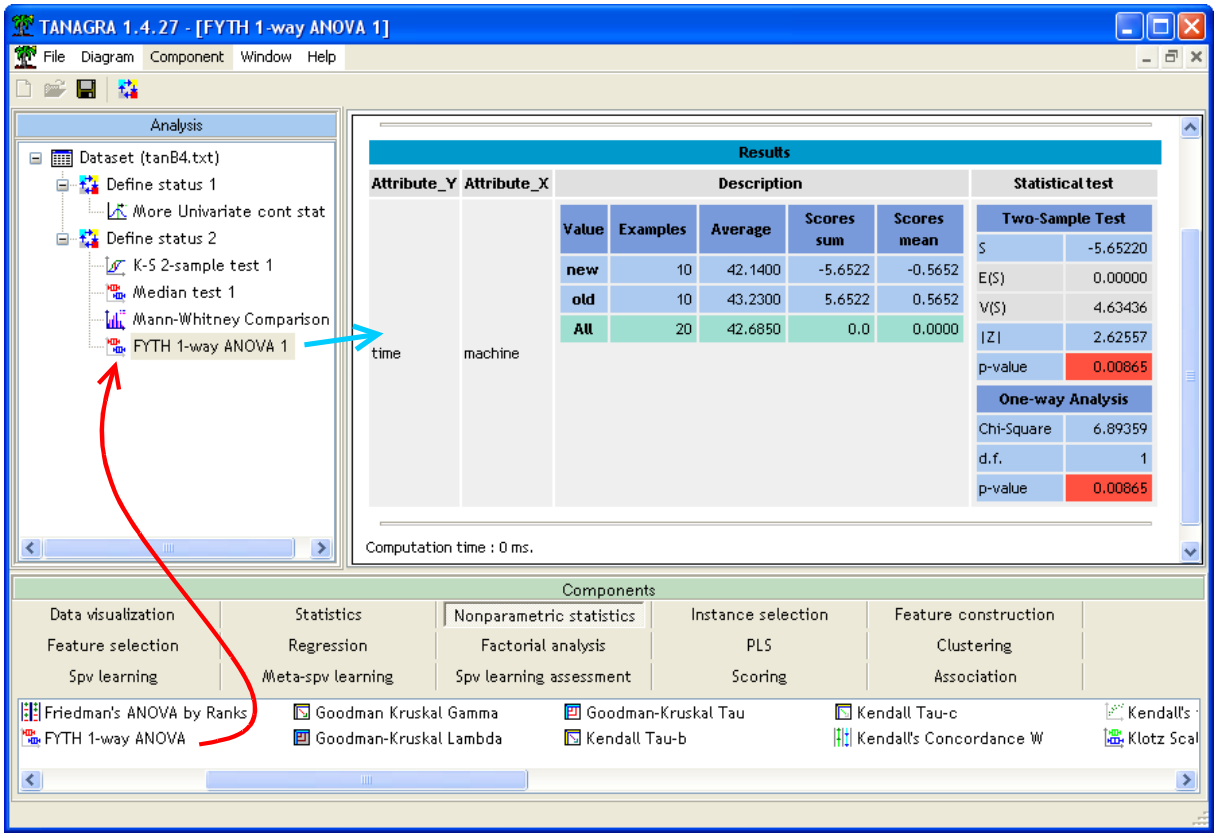
5.3 Autres tests non paramétriques pour un modèle de localisation

Les tests de Fisher-Yates-Terry-Hoeffding (FYTH) et de Van der Waerden sont des variantes de Wilcoxon-Mann-Whitney. On les met usuellement en avant lorsqu'il s'agit de traiter des variables

[4] http://www.stat.psu.edu/online/development/stat500_spss/lesson10/lesson10_03.html

dont la distribution se rapproche de la loi de Gauss. La préparation des données est réalisée en 2 temps : (1) elles sont tout d'abord transformées en rangs ; (2) ces derniers sont transformés en **scores normaux** (ou codes normaux) en utilisant les quantiles de la loi normale. La statistique de tests correspond alors à la somme des scores.

Nous insérons le composant FYTH 1-WAY ANOVA (onglet NONPARAMETRIC STATISTICS) dans le diagramme. Nous obtenons le résultat suivant.

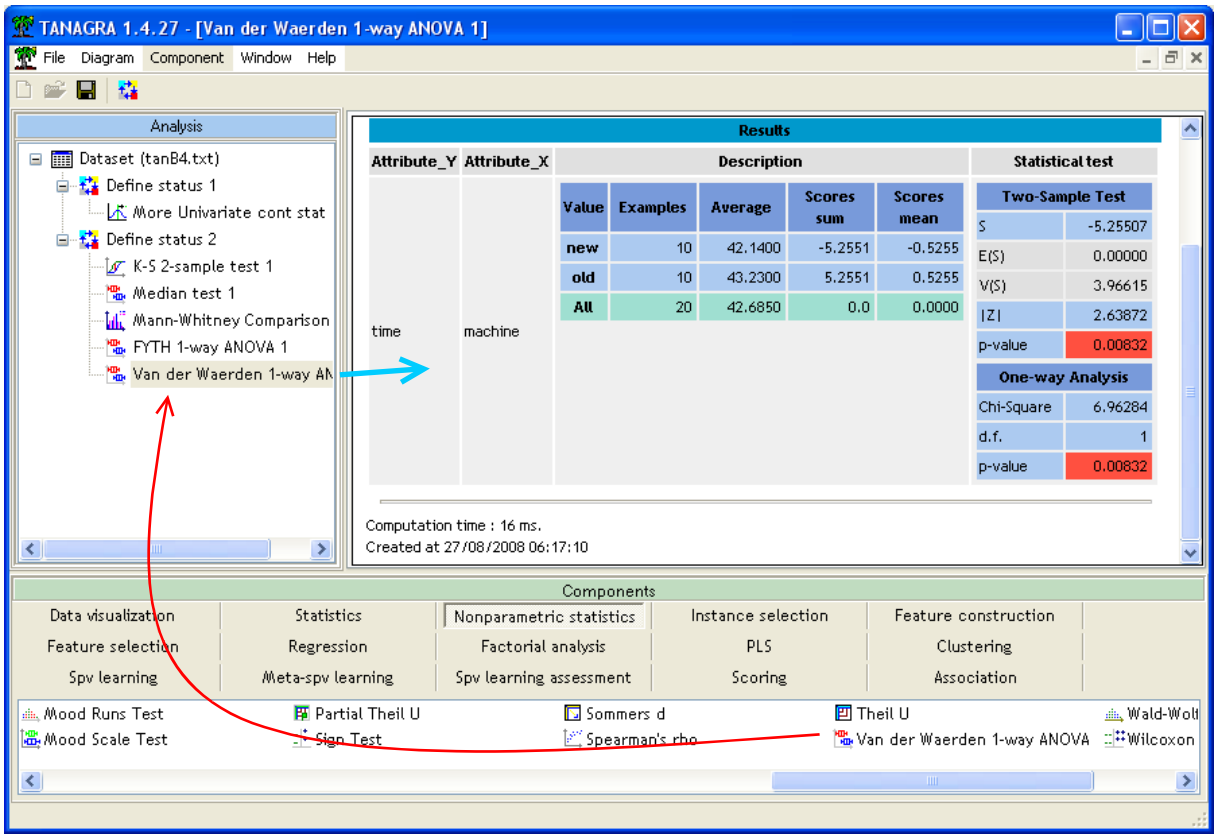


La statistique du test, la somme des scores du groupe de référence, est S = -5.6522. Sous H0, son espérance et sa variance sont respectivement E(S) = 0 et V(S) = 4.63436. Nous pouvons former la statistique standardisée

$$|Z| = \frac{|S - E(S)|}{\sqrt{V(S)}} = \frac{|-5.6522|}{\sqrt{4.63436}} = 2.62557$$

Avec une probabilité critique p = 0.00865. Les résultats sont très proches de ceux de Wilcoxon-Mann-Whitney.

Le test de Van der Waerden diffère très légèrement de FYTH par le mode de calcul du quantile. Nous introduisons le composant VAN DER WAERDEN 1-WAY ANOVA (onglet NONPARAMETRIC STATISTICS) dans le diagramme. Les conclusions sont très similaires à celles de FYTH.



6 Conclusion

Les composants MEDIAN TEST, FYTH 1-WAY ANOVA et VAN DER WAERDEN sont opérationnels même si la variable INPUT catégorielle possède (K > 2) modalités. Dans ce cas, la statistique Z est masquée, seul le résultat associé à l'analyse de variance sur les scores (χ^2) est proposé.

Cela n'est pas possible pour le composant MANN-WHITNEY COMPARISON. On lui substituera alors le test de Kruskal-Wallis (composant KRUSKAL-WALLIS 1-WAY ANOVA) qui est son alter ego pour la comparaison de (K > 2) populations.