1 Objectif

Comparaison de populations. Tests paramétriques multivariés avec Tanagra.

Les **tests de comparaison de populations** visent à déterminer si K (K ≥ 2) échantillons proviennent de la même population au regard d'une groupe de variables d'intérêt (X1,...,Xp). 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.

On parle de tests **paramétriques** lorsque l'on fait l'hypothèse que X suit une distribution paramétrée. Dès lors comparer les distributions empiriques conditionnelles revient à comparer les paramètres : la moyenne et la variance lorsque l'on fait l'hypothèse de normalité en analyse univariée ; le vecteur moyenne et la matrice de variance covariance lorsque l'on considère que le groupe de variables est **distribuée selon une loi normale multidimensionnelle** en analyse multivariée.

Enfin, dans ce didacticiel, nous traitons les tests **multivariés** c.-à-d. nous **étudions simultanément plusieurs variables d'intérêt**.

Ce type de test peut servir à comparer effectivement des processus (ex. est-ce que deux machines produisent des boulons de même diamètre et qualité), mais il permet également d'éprouver la liaison qui peut exister entre une variable catégorielle et une variable quantitative (ex. est ce que les femmes conduisent en moyenne moins vite que les hommes, provoquent moins d'accidents et consomment moins ?).

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_Parametriques.pdf (Partie III). Les tests d'écrits dans ce didacticiel s'appliquent aux échantillons indépendants. Les procédures pour échantillons appariés feront l'objet d'autres didacticiels.

2 Données

Le fichier CREDIT_APPROVAL.XLS[1] décrit 50 ménages, formés de couples mariés, tous deux actifs, qui ont déposé une demande de crédit auprès d'un établissement bancaire. Les variables disponibles sont les suivantes :

Variable	Description
Sal.Homme	Logarithme du salaire de l'homme
Sal.Femme	Logarithme du salaire de la femme
Rev.Tete	Logarithme du revenu par tête c.-à-d. total des revenus divisé par le nombre de personnes
Age	Logarithme de l'âge de l'homme
Acceptation	Accord du crédit par l'organisme prêteur
Garantie.Supp	Garantie supplémentaire demandée par l'organisme prêteur
Emploi	Type d'emploi occupé par la personne de référence lors de la demande de crédit

Les variables quantitatives sont les variables d'intérêt ; les variables catégorielles vont servir à définir les sous populations.

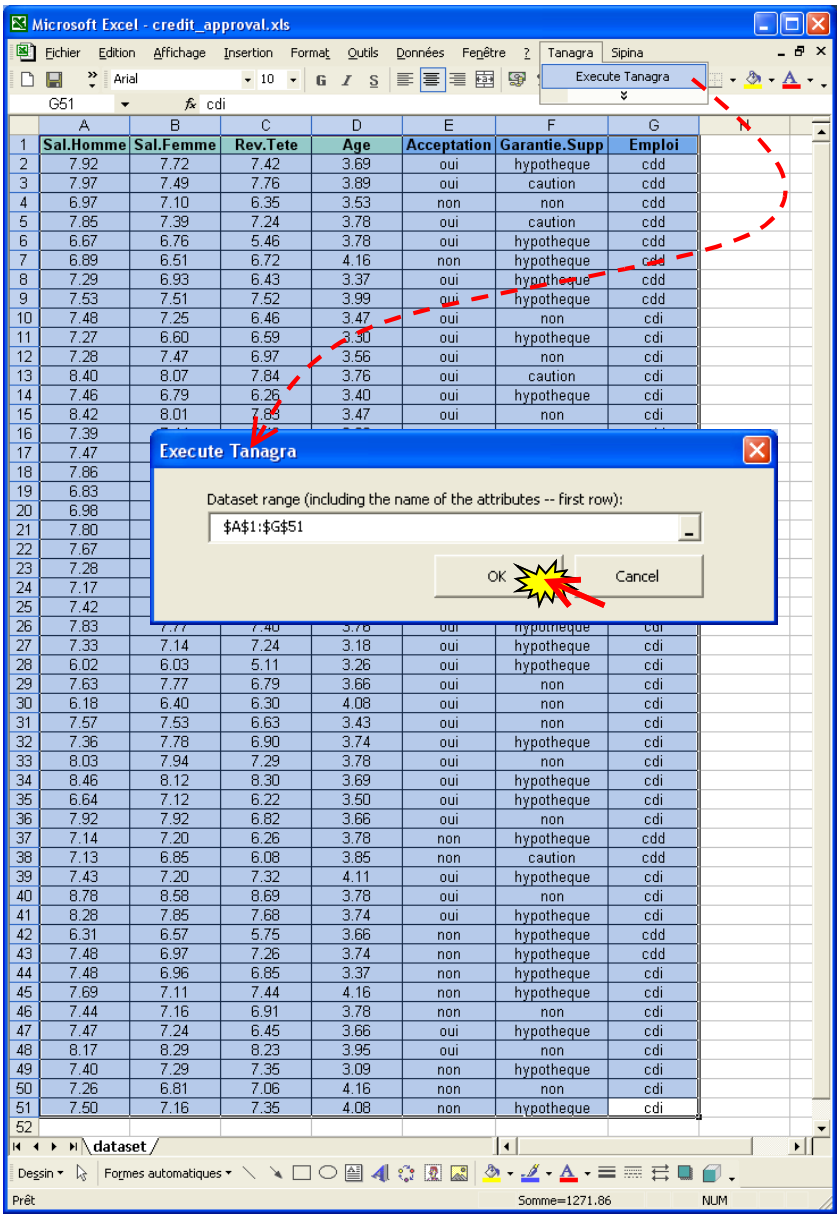
[1] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/credit_approval.xls

3 Comparaison de 2 moyennes – T² de Hotelling

« Moyenne » s'entend barycentre du nuage de points. Il s'agit d'un vecteur de dimension (p x 1) où p est le nombre de variables d'intérêt. A la composante n°j correspond la moyenne de la j^{ème} variable. Dans notre ensemble de données, p = 4.

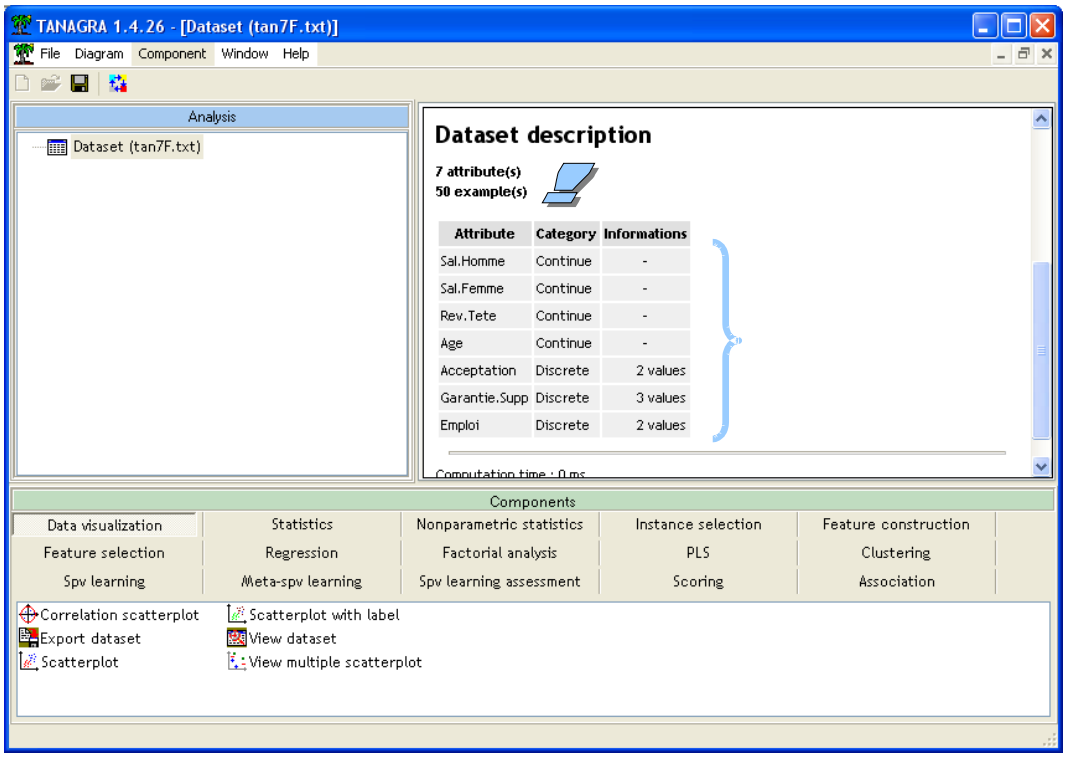
3.1 Importer les données dans Tanagra

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



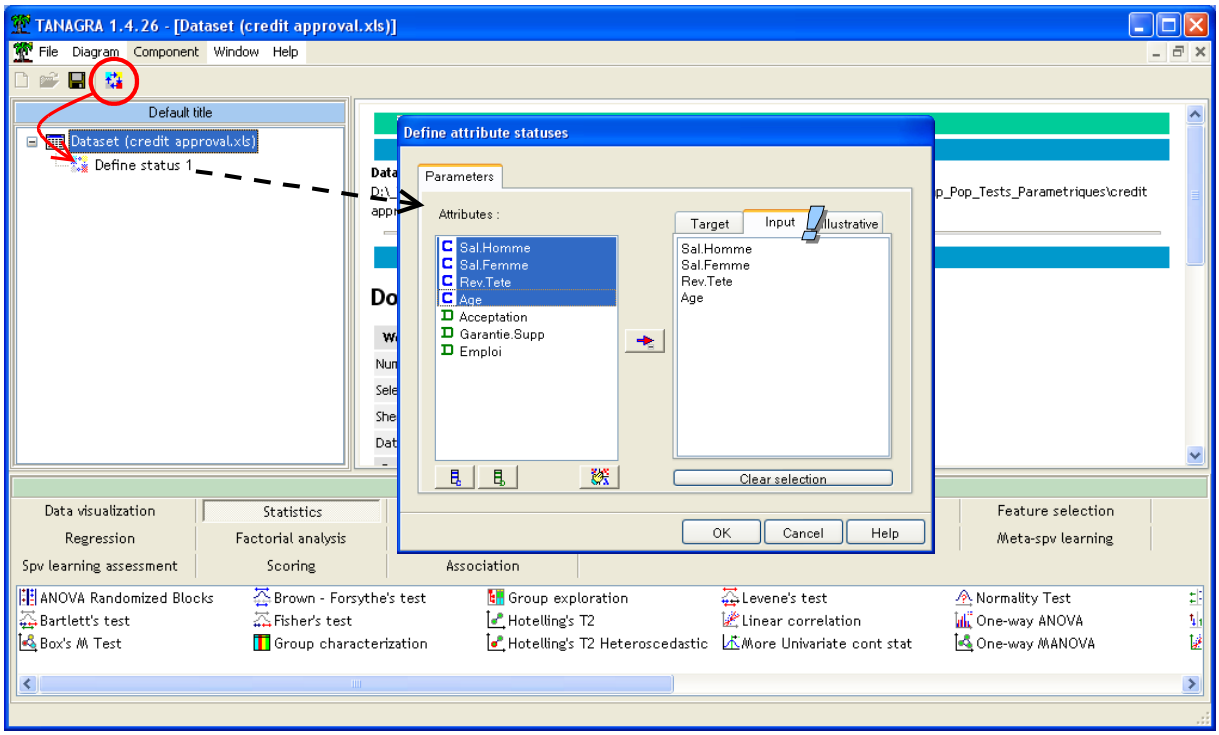
[2] Voir http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-macro.html concernant l'installation et l'utilisation de la macro complémentaire TANAGRA.XLA.

Tanagra est automatiquement lancé. Un nouveau diagramme est créé. Nous vérifions que l'ensemble de données comporte 50 observations et 7 variables.

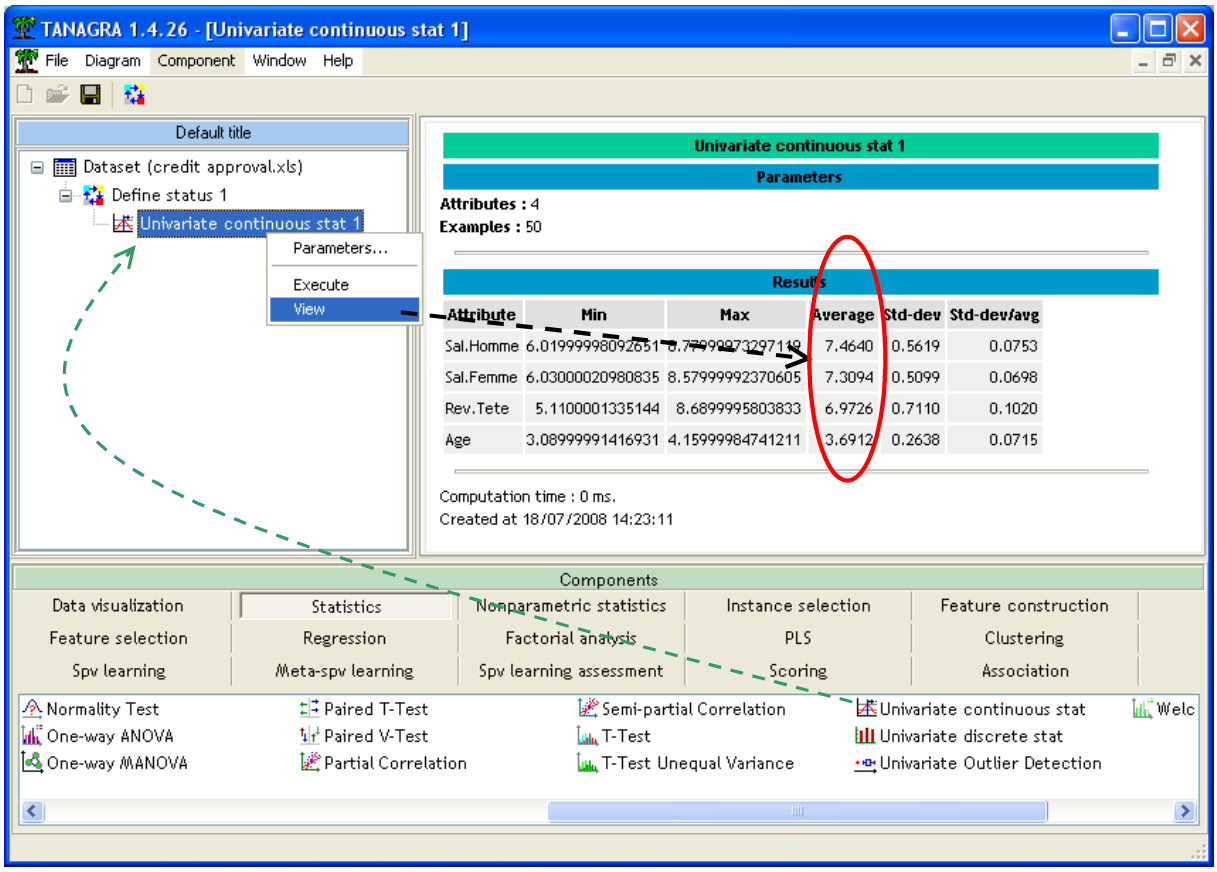


3.2 Statistiques descriptives

Première étape toujours, décrivons un peu nos variables d'intérêt. Nous insérons le composant DEFINE STATUS dans le diagramme via le raccourci dans la barre d'outils. Nous plaçons les variables continues en INPUT.



Nous insérons ensuite le composant UNIVARIATE CONTINUOUS STAT (onglet STATISTICS). Nous activons le menu contextuel VIEW.



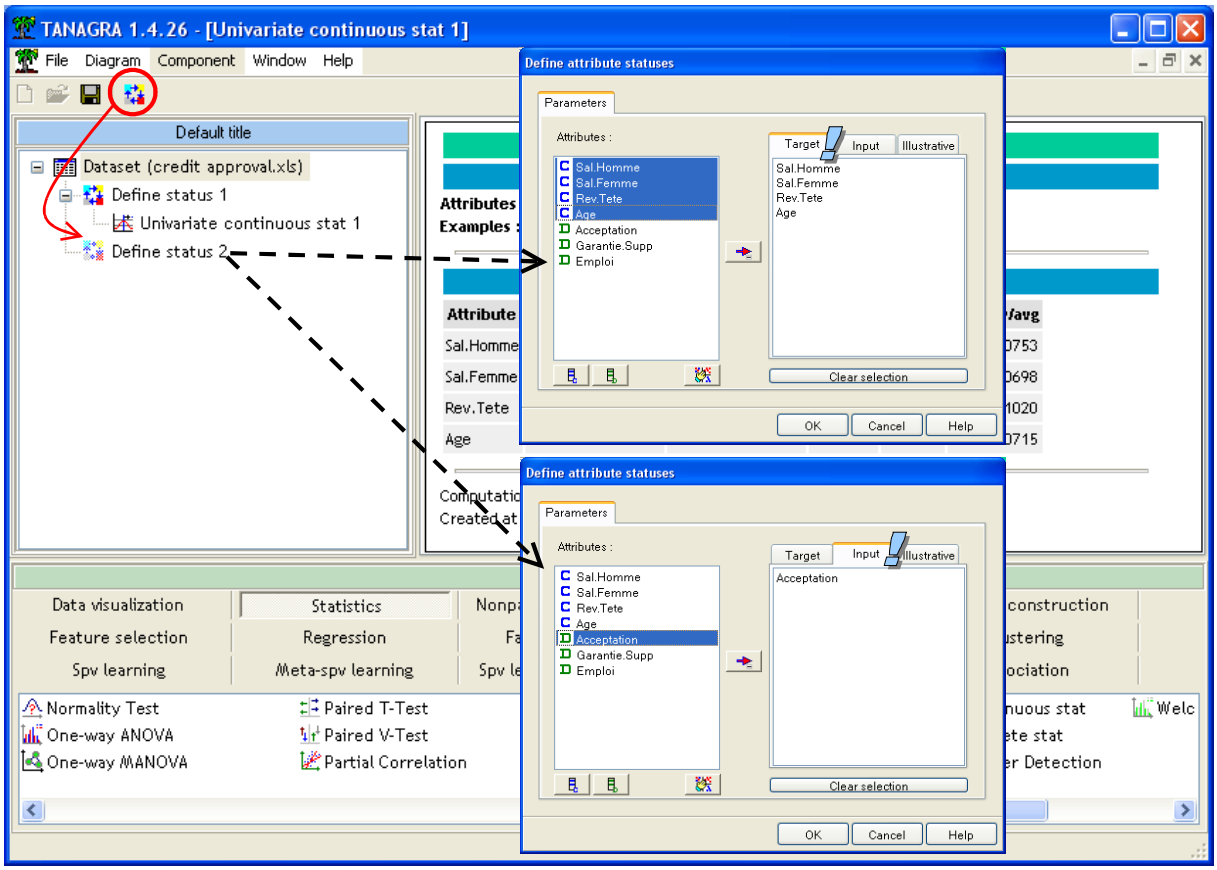
La colonne AVERAGE nous intéresse particulièrement dans cette section. Nous y lisons le barycentre du vecteur moyenne calculé sur l'ensemble des observations $\bar{X}' = (7.4640; 7.3094; 6.9726; 3.6912)$.

L'objectif de la comparaison de moyennes est de vérifier que ce vecteur prend des valeurs significativement différentes dans les sous populations que l'on souhaite confronter.

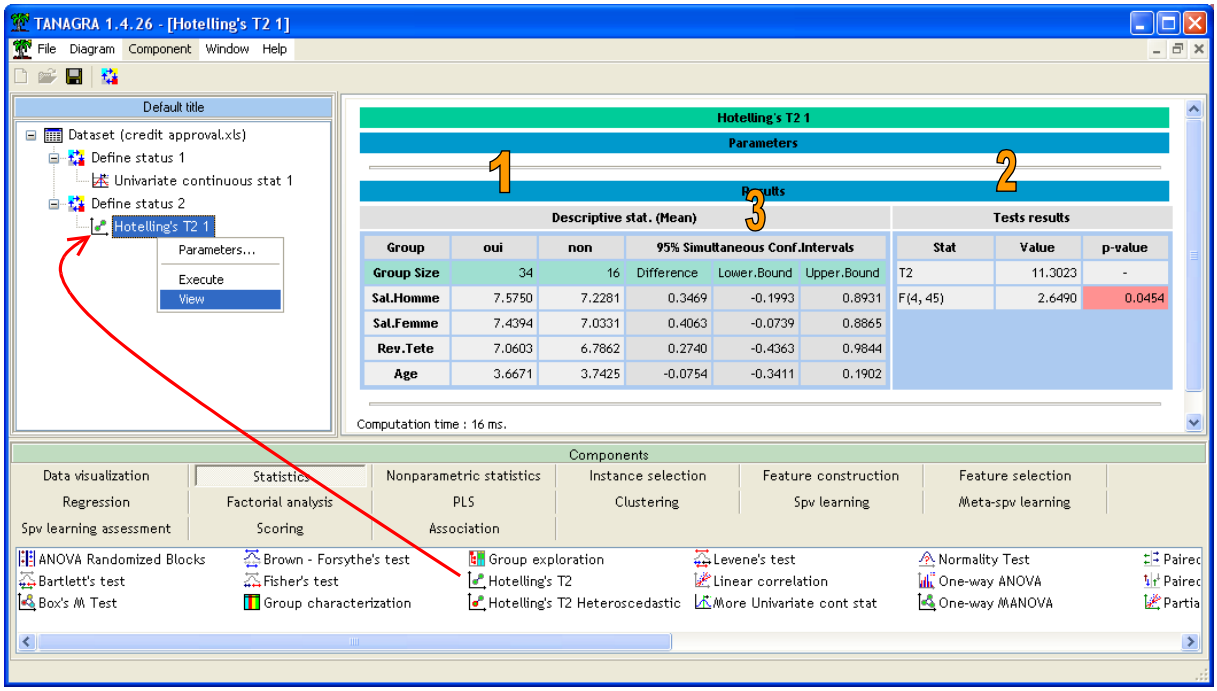
3.3 T² de Hotelling sous l'hypothèse d'homoscédasticité

Ce test vise à comparer 2 vecteurs de moyennes. Il suppose que les dispersions dans les sous populations sont identiques c.-à-d. les matrices de variance covariance conditionnelles sont les mêmes. Une matrice commune sera calculée, la matrice de variance covariance intra classes.

Dans notre exemple, nous souhaitons comparer les caractéristiques des clients selon la décision de la banque concernant l'octroi du crédit. Nous introduisons DEFINE STATUS dans le diagramme, nous plaçons en TARGET les variables d'intérêt, en INPUT la variable ACCEPTATION.



Nous insérons ensuite le composant HOTELLING'S T2 dans le diagramme. Nous activons le menu VIEW pour accéder aux résultats.



Plusieurs éléments doivent retenir notre attention :

- (1) Dans le tableau des statistiques descriptives, nous observons les vecteurs des moyennes conditionnelles, relatives aux groupes ACCEPTATION = OUI et ACCEPTATION = NON.

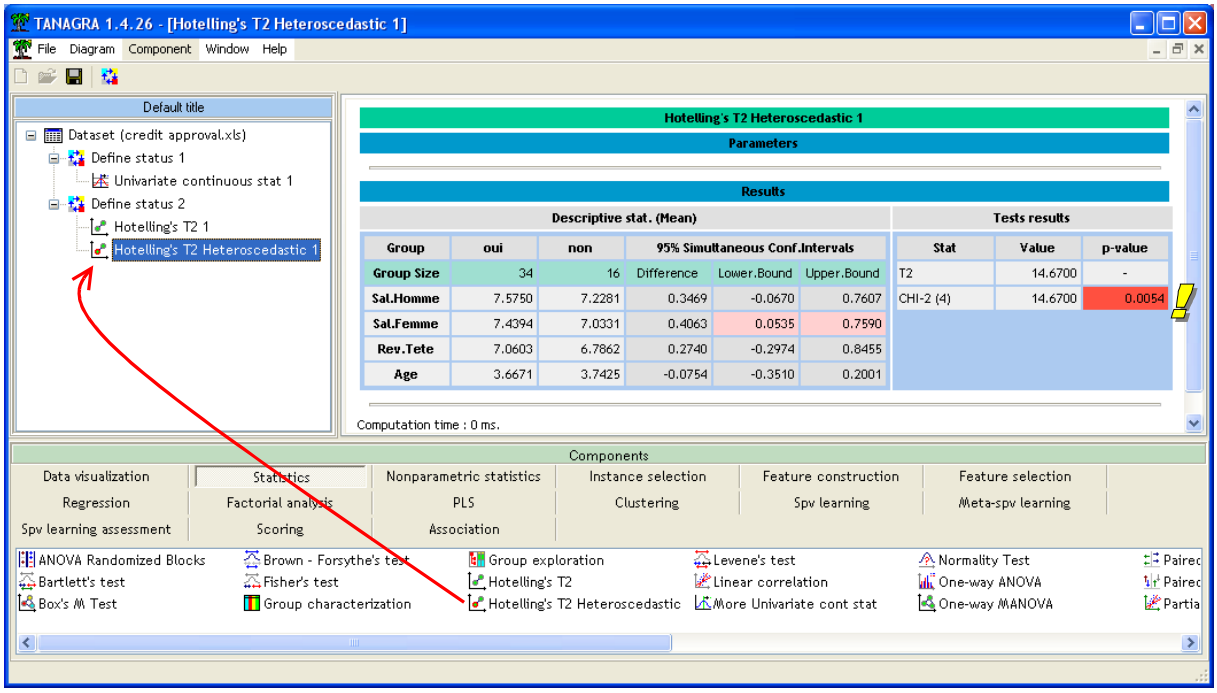
- (2) Dans le dernier tableau à droite, nous obtenons la statistique du test T^2, et sa transformation distribuée selon la loi de Fisher à (4 ; 45) degrés de liberté. Au risque 5%, nous rejetons l'hypothèse nulle d'égalité des vecteurs.

- (3) Revenons dans le tableau des statistiques descriptives. Dans les 3 dernières colonnes nous avons le vecteur des écarts entre les moyennes, et leurs intervalles de confiance simultanés au niveau 95%. Ces indications nous permettent de savoir si l'écart est significatif à 5% sur une des variables d'intérêt que l'on souhaite, *a priori*, étudier en particulier[3]. Il l'est si l'intervalle ne contient pas la valeur **0**. Dans notre exemple, il semble que l'écart global significatif détecté par le T^2 ne soit pas spécialement associé à une des variables.

3.4 T^2 de Hotelling sous l'hypothèse d'hétéroscédasticité

Si l'hypothèse d'égalité des matrices de variance covariance est discutable, surtout lorsque les effectifs sont déséquilibrés, nous avons intérêt à utiliser la variante qui introduit explicitement le calcul différencié des matrices conditionnelles.

Toujours à la suite du DEFINE STATUS précédent, nous insérons le composant HOTELLING'S T2 HETEROSCEDASTIC (onglet STATISTICS).



La statistique $T^2 = 14.6700$ suit asymptotiquement une loi du KHI-2 à p degrés de liberté[4]. L'écart entre les vecteurs de moyennes est largement significatif à 5% (p-value = 0.0054). Et on constate maintenant qu'elle serait plus particulièrement imputable au salaire féminin du couple. L'intervalle de confiance afférente ne contient pas la valeur **0**.

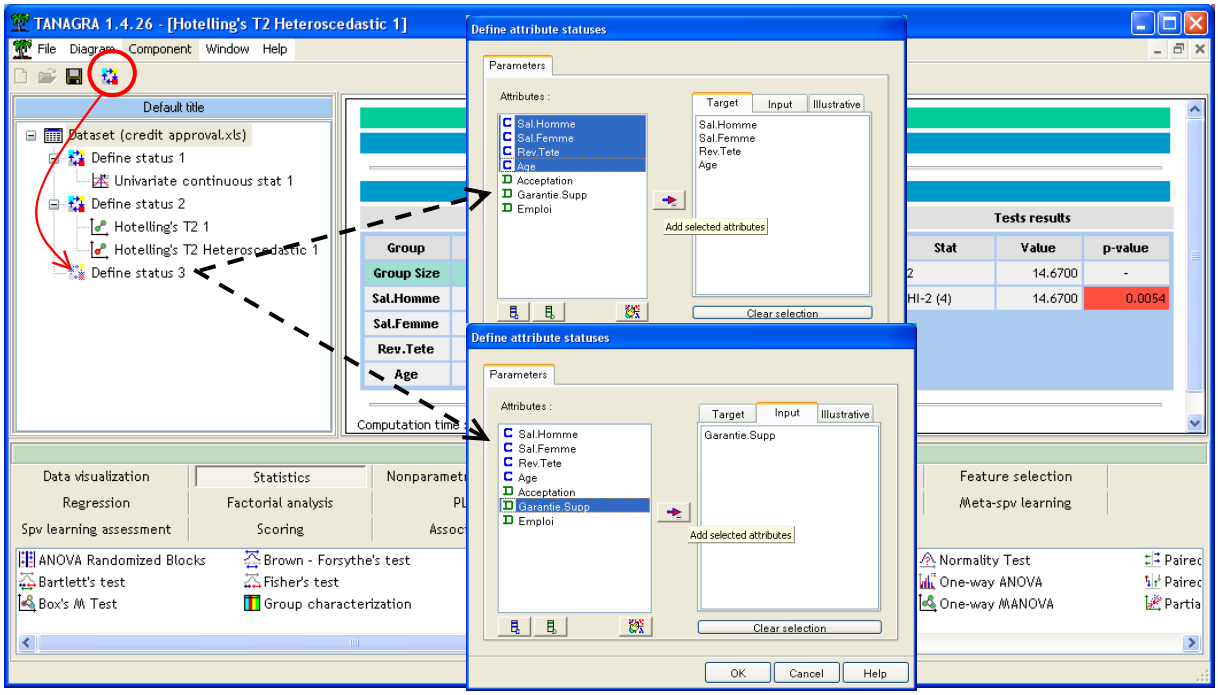
[3] http://www.stat.psu.edu/online/development/stat505/10_2sampHotel/05_2sampHotel_differ.html

[4] Il existe une approximation plus précise qui suit une loi de Fisher, préférable pour les petits effectifs. Elle n'est pas disponible pour l'instant dans Tanagra (version 1.4.26).

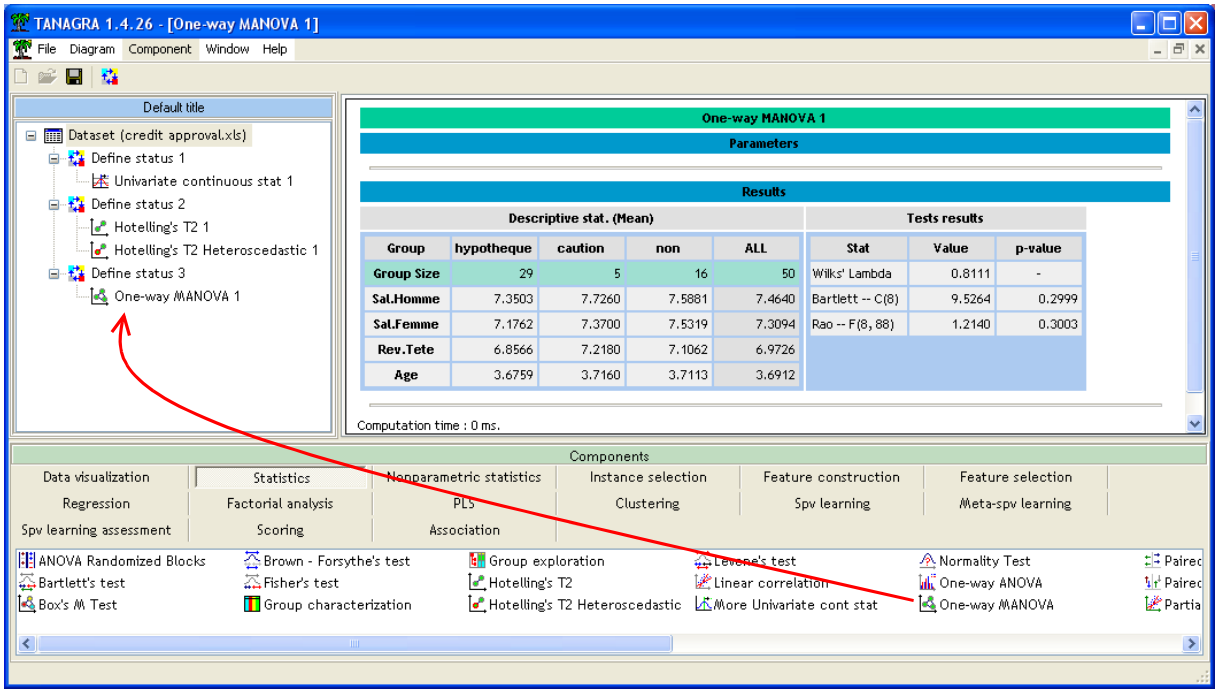
4 Comparaison de K moyennes – Lambda de Wilks

L'objectif maintenant est de comparer les vecteurs de moyennes selon le type de garantie adopté par les demandeurs de crédit (K = 3 modalités). Nous sommes dans le canevas de l'analyse de variance multivariée à un facteur (MANOVA).

Nous insérons un nouveau DEFINE STATUS dans le diagramme. Nous plaçons en TARGET nos variables d'intérêt, en INPUT la variable GARANTIE.SUPP.



Nous ajoutons le composant ONE-WAY MANOVA (onglet STATISTICS) dans le diagramme.



Dans la partie DESCRIPTIVE STAT., nous observons les vecteurs des moyennes conditionnelles pour les 3 groupes, et la moyenne globale pour l'ensemble des observations (ALL ; à rapprocher avec les résultats de la section 3.2).

La statistique LAMBDA de WILKS = 0.8111. Plus elle se rapproche de la valeur **1**, moins les écarts sont significatifs. Deux transformations sont disponibles, celle de Bartlett, suffisante pour les grands effectifs, et celle de Rao, recommandée pour les petits effectifs. Dans les deux cas, elles aboutissent à la même conclusion : à 5%, les données disponibles ne permettent pas de rejeter l'hypothèse d'égalité des moyennes.

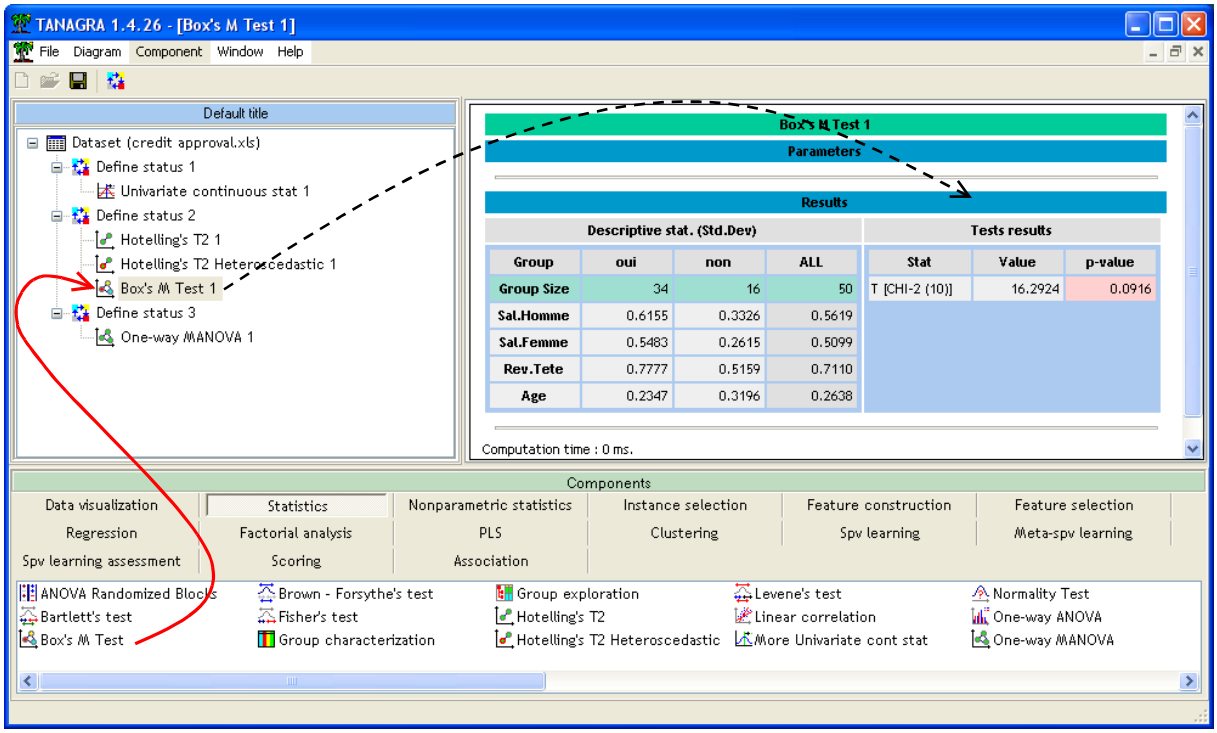
5 Comparaison de K variances – La statistique M de Box

Le test de Hotelling (section 3.3) et le test de Wilks (section 4) supposent que les matrices de variance covariance sont identiques dans les sous populations. Il est possible d'éprouver cette hypothèse à l'aide de la version multidimensionnelle du test de Bartlett, que l'on retrouve également sous l'appellation « Test M de Box » dans la littérature.

Nous allons mettre en œuvre cette procédure pour les comparaisons de moyennes réalisées précédemment.

5.1 Vérification pour le test de Hotelling

Nous insérons le composant BOX'S M TEST (onglet STATISTICS) en dessous de DEFINE STATUS 2, au même niveau que les tests de Hotelling. Nous activons le menu VIEW.

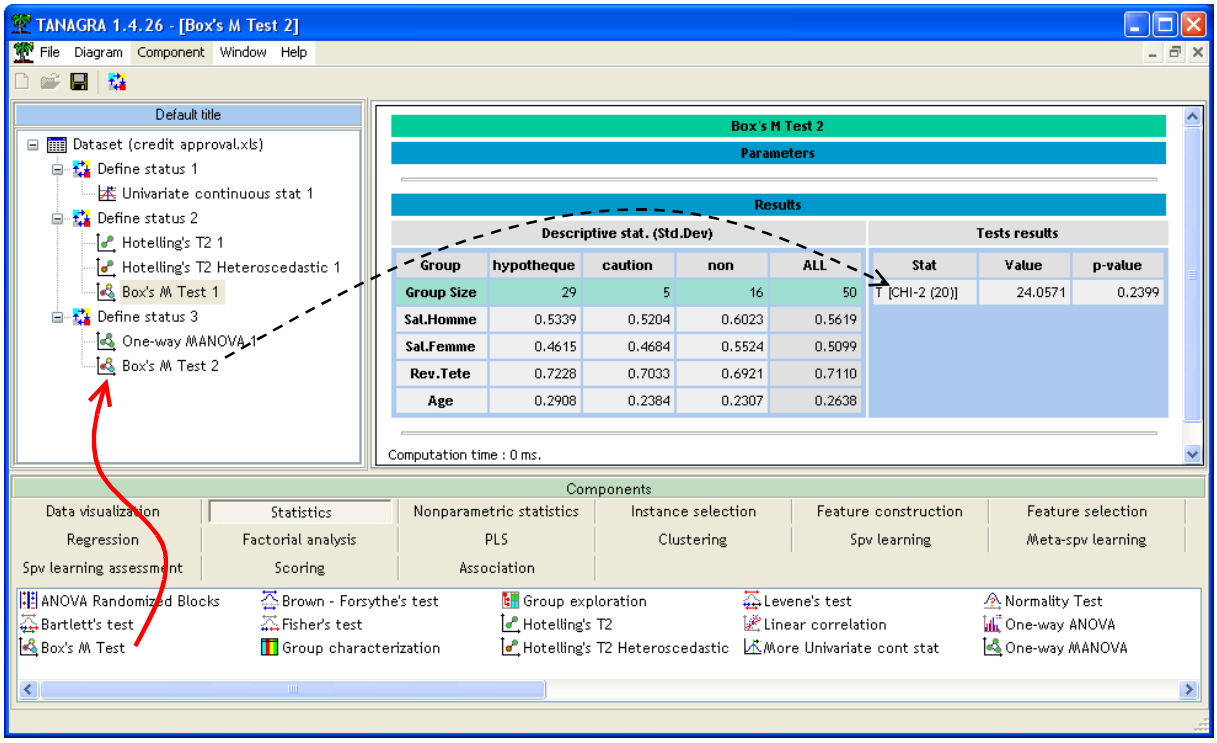


L'écart type marginal et les écarts types conditionnels sont fournies dans les statistiques descriptives. Ces valeurs sont données à titre indicatif. Elles donnent une idée sur les éventuelles différences, de manière univariée, sur chaque variable d'intérêt. Il ne saurait question de tirer des conclusions à partir de ces statistiques descriptives, il manque de toute manière les informations sur les covariances.

Le test de comparaison des matrices de variance covariance propose la statistique T = 16.2924. Elle suit une loi du KHI-2 à 10 degrés de liberté. Les matrices ne sont pas significativement différentes à 5%, elles le sont en revanche à 10% (p-value = 0.0916). Il semble, *a posteriori*, que le test de Hotelling introduisant une estimation différenciée des matrices de variance covariance soit le plus approprié lors de la comparaison de moyennes, conviction renforcée par l'effectif relativement faible et déséquilibré des groupes.

5.2 Vérification pour le test de Wilks

Nous réalisons la même vérification pour la MANOVA. Nous insérons de nouveau le composant BOX'S M TEST.



Ici en revanche, l'hypothèse d'égalité des matrices de variance covariance conditionnelles est tout à fait crédible avec un T = 24.0571, et surtout une p-value = 0.2399.

A posteriori, les résultats associés au Lambda de Wilks lors de la comparaison des K vecteurs de moyennes sont confirmés.