1 Objectif

Comparaison de populations. Tests paramétriques univariés avec Tanagra.

Les **tests de comparaison de populations** visent à déterminer si K ($K \geq 2$) 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.

On parle de tests **paramétriques** lorsque l'on fait l'hypothèse que la variable X suit une distribution paramétrée. Dès lors comparer les distributions empiriques conditionnelles revient à comparer les paramètres, soit la moyenne et la variance lorsque l'on fait l'hypothèse de normalité de X.

Enfin, dans ce didacticiel, nous traitons les tests **univariés** c.-à-d. nous étudions une seule variable d'intérêt. Lorsque nous traitons simultanément plusieurs variables, on parle de tests multivariés. Ce qui fera l'objet d'un autre didacticiel prochainement.

Ce type de test[1] peut servir à comparer effectivement des processus (ex. est-ce que deux machines produisent des boulons de même diamètre), 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 sur telle portion de route).

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 (Parties I et II). Nous utiliserons les mêmes données et nous suivrons exactement la même trame pour que le lecteur puisse suivre le détail des formules mises en œuvre.

2 Données

Le fichier CREDIT_APPROVAL.XLS[2] 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 toutes potentiellement « variable d'intérêt » ; les variables catégorielles vont servir à définir les sous populations.

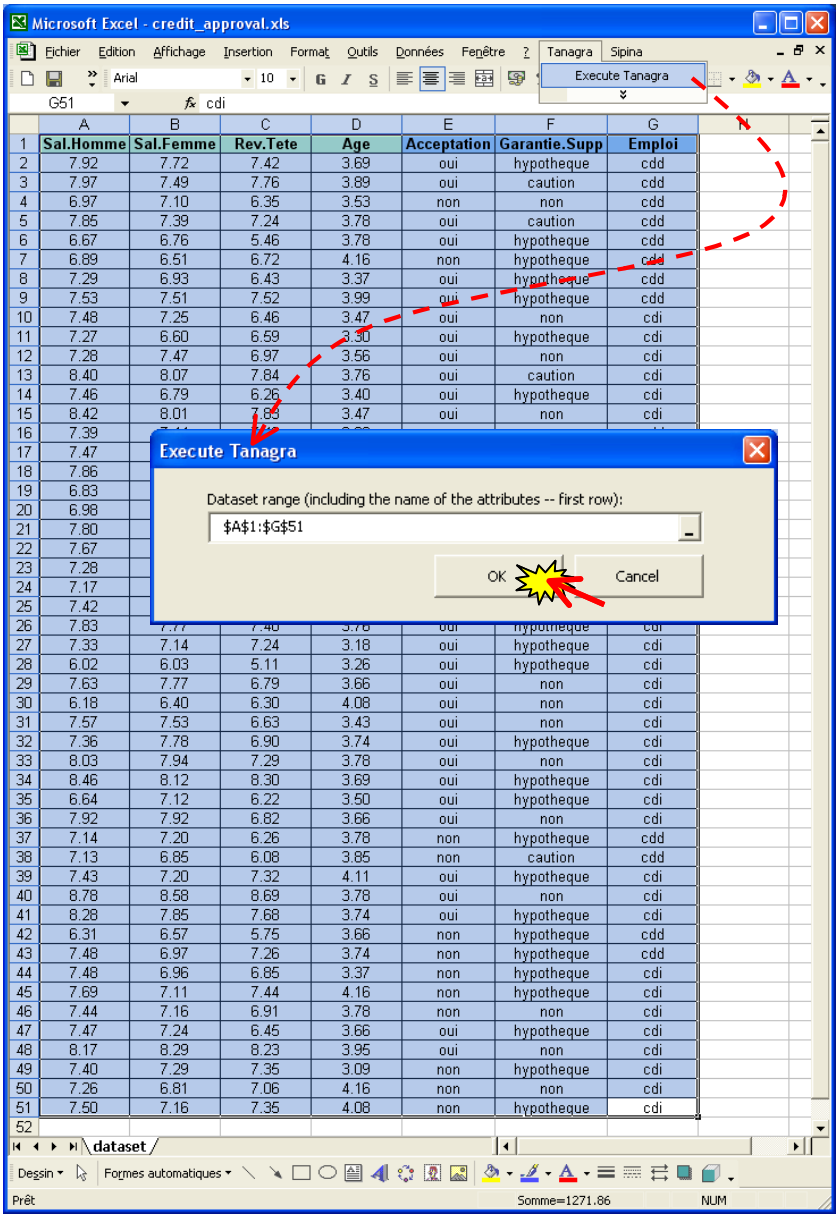
[1] http://www.itl.nist.gov/div898/handbook/prc/prc.htm

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

3 Statistiques descriptives et test de normalité

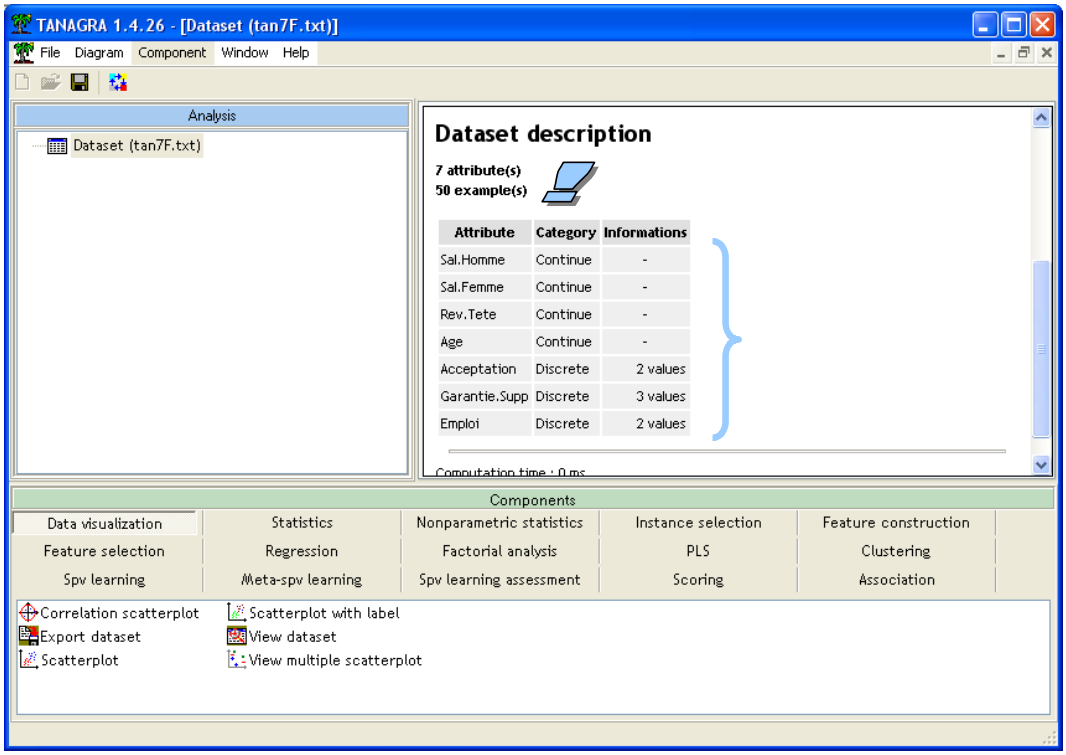
3.1 Importer les données dans Tanagra

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



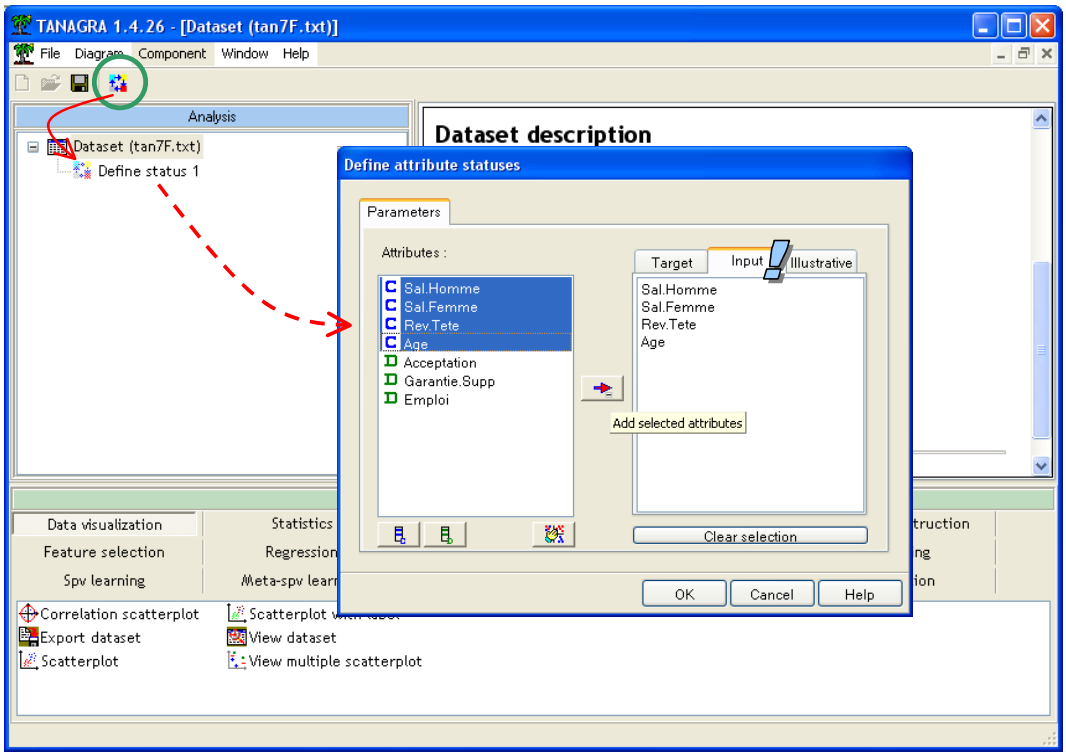
Tanagra est automatiquement lancé. Un nouveau diagramme est créé. Nous vérifions que l'ensemble de données comporte 50 observations et 7 variables.

[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.

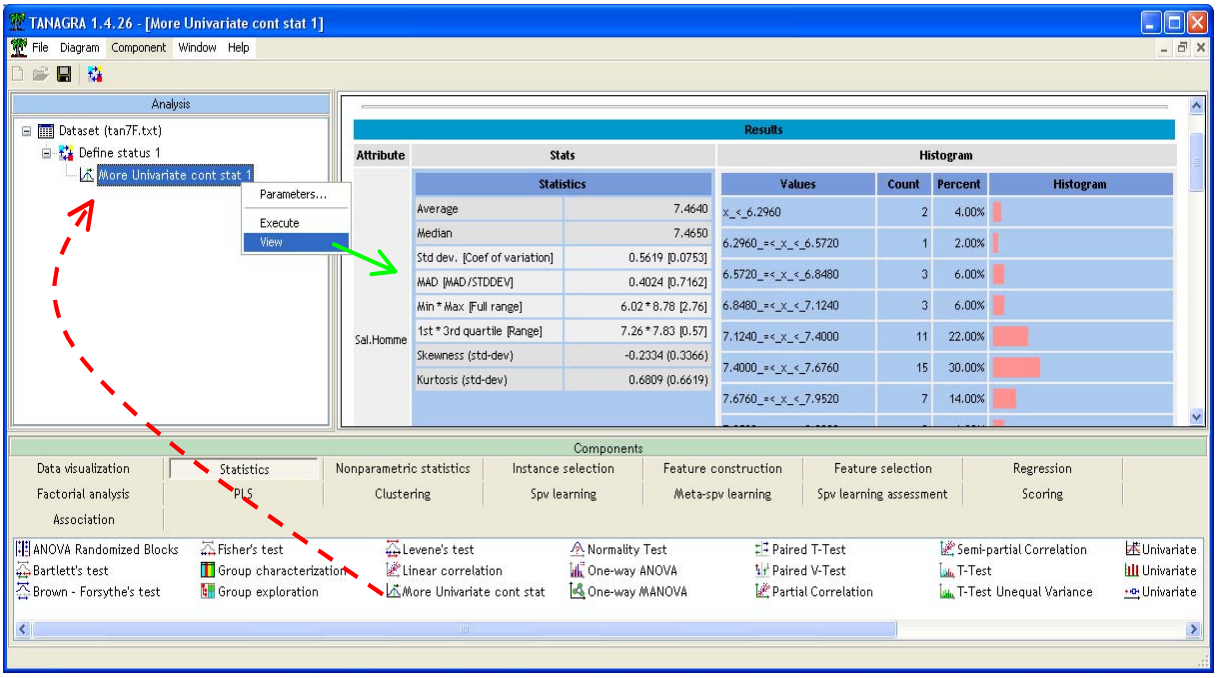


3.2 Statistiques descriptives

Première étape, nous souhaitons inspecter les variables quantitatives à l'aide d'indicateurs statistiques simples. Nous insérons le composant DEFINE STATUS dans le diagramme, via le raccourci dans la barre d'outils, puis nous plaçons en INPUT les variables Sal.Homme, Sal.Femme, Rev.Tete et Age.



Nous introduisons alors le composant MORE UNIVARIATE CONT STAT (onglet STATISTICS). Nous cliquons sur VIEW pour obtenir les résultats.



Avec ce type d'outils, on cherche avant tout à détecter les anomalies fortes du type « distribution asymétrique », « bimodale », l'existence de points atypiques, etc. Il semble qu'il n'y ait pas lieu de à s'inquiéter outre mesure ici.

3.3 Test de normalité

Les tests que nous présentons dans ce didacticiel supposent gaussienne la distribution des variables d'intérêt. Certes, ils sont plus ou moins robustes face à cette hypothèse. Néanmoins, pour assurer l'affaire (et montrer comment faire), nous décidons de procéder à cette vérification[4,5].

Une première évaluation empirique était déjà possible avec le composant précédent. On sait que le rapport entre l'écart absolu moyen (MAD : *mean absolute deviation*) et l'écart type (STD-DEV : *standard deviation*) est approximativement égal à 0.8 lorsque la distribution est gaussienne[6]. Attention, l'inverse n'est pas vrai. Ce n'est pas parce que le ratio est proche de 0.8 que la distribution est forcément normale. On utilisera surtout cette règle empirique pour détecter les écarts évidents, annonciateurs de problèmes. Dans notre cas, il est n'y a pas d'écarts manifestes. En effet, il est égal à 0.7162 pour « Sal.Homme », 0.7846 pour « Sal.Femme », 0.7986 pour « Rev.Tete » et 0.7895 pour « Age ». Mis à part « Sal.Homme », nous sommes même étonnamment proches de la valeur de référence.

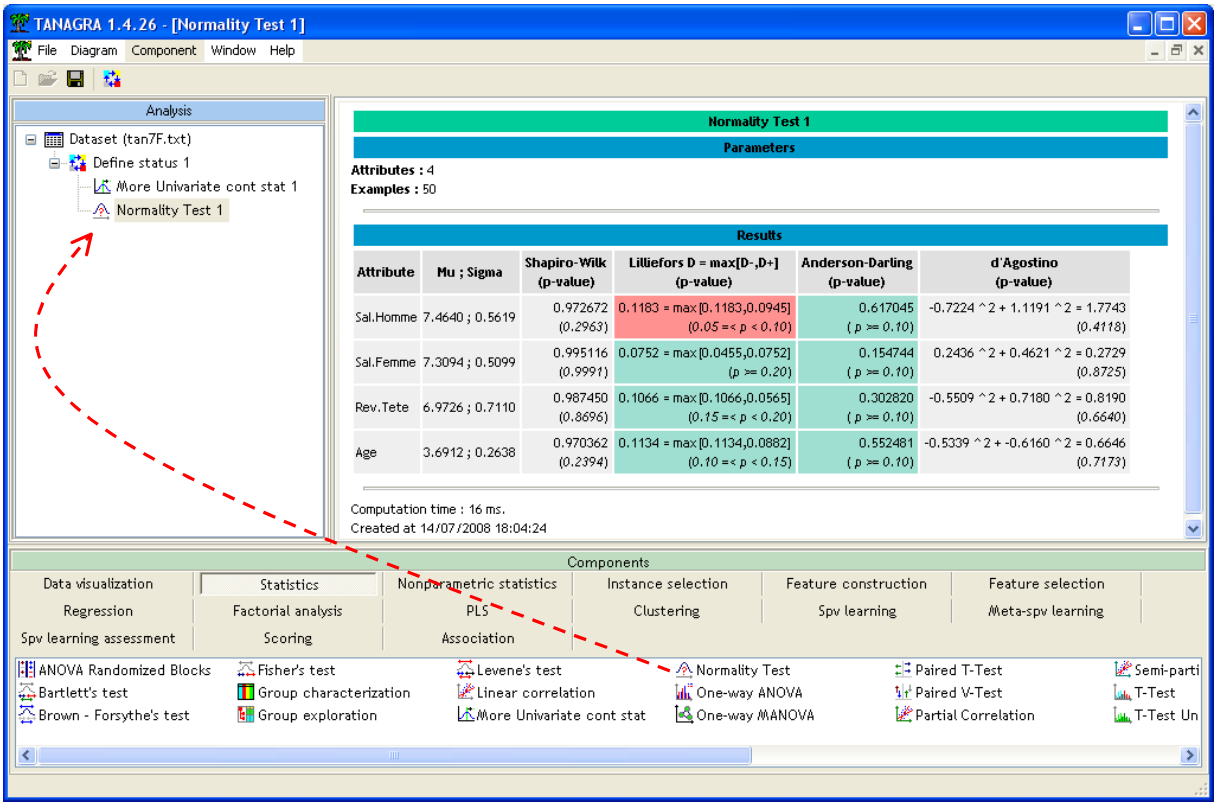
Procédons maintenant à une évaluation plus formelle en utilisant les tests d'adéquation à la loi normale. Nous insérons le composant NORMALITY TEST (onglet STATISTICS) dans le diagramme. Nous cliquons sur le menu contextuel VIEW.

---

[4] Voir http://tutoriels-data-mining.blogspot.com/2008/04/tests-dadquation-la-loi-normale.html pour plus de détails sur les tests d'adéquation à la loi normale.

[5] **Attention** : Le test de normalité sur la globalité de l'échantillon n'est pas forcément une bonne idée. En effet, prenons l'exemple d'une comparaison de moyennes sur échantillons indépendants. Si l'hypothèse alternative est vraie, la distribution des données regroupées ne peut pas être compatible avec l'hypothèse de normalité puisqu'elle sera bimodale. Il est souvent plus indiqué de réaliser les tests de normalité à l'intérieur des sous populations.

[6] $\sqrt{\frac{2}{\pi}} \approx 0.798$ pour être précis. Voir http://en.wikipedia.org/wiki/Mean_absolute_deviation



Au risque 5%, l'hypothèse de compatibilité des variables (prises individuellement) avec la loi normale n'est pas remise en cause par les données. Ceci n'est guère étonnant, nous avons procédé à des ajustements, en passant par les logarithmes principalement, de manière à rendre au moins symétriques les distributions.

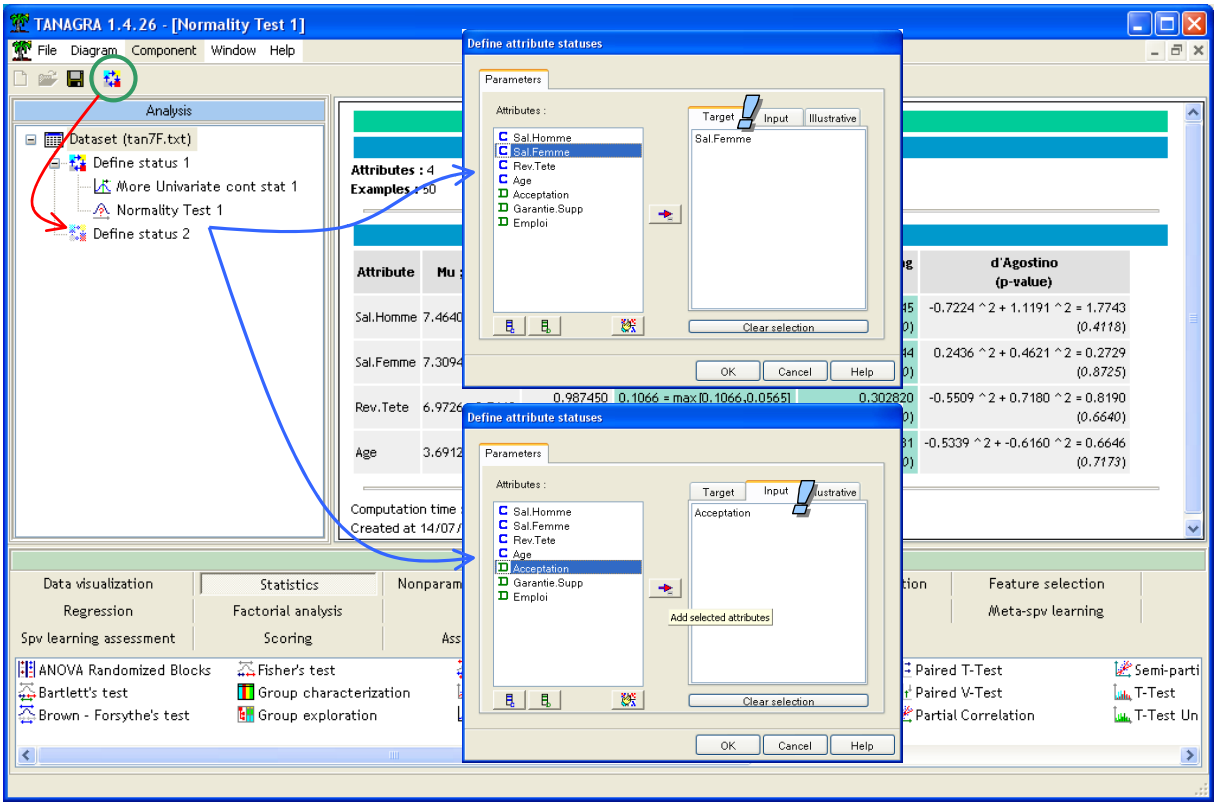
4 Tests pour échantillons indépendants

4.1 Comparaison de 2 moyennes – Variances égales

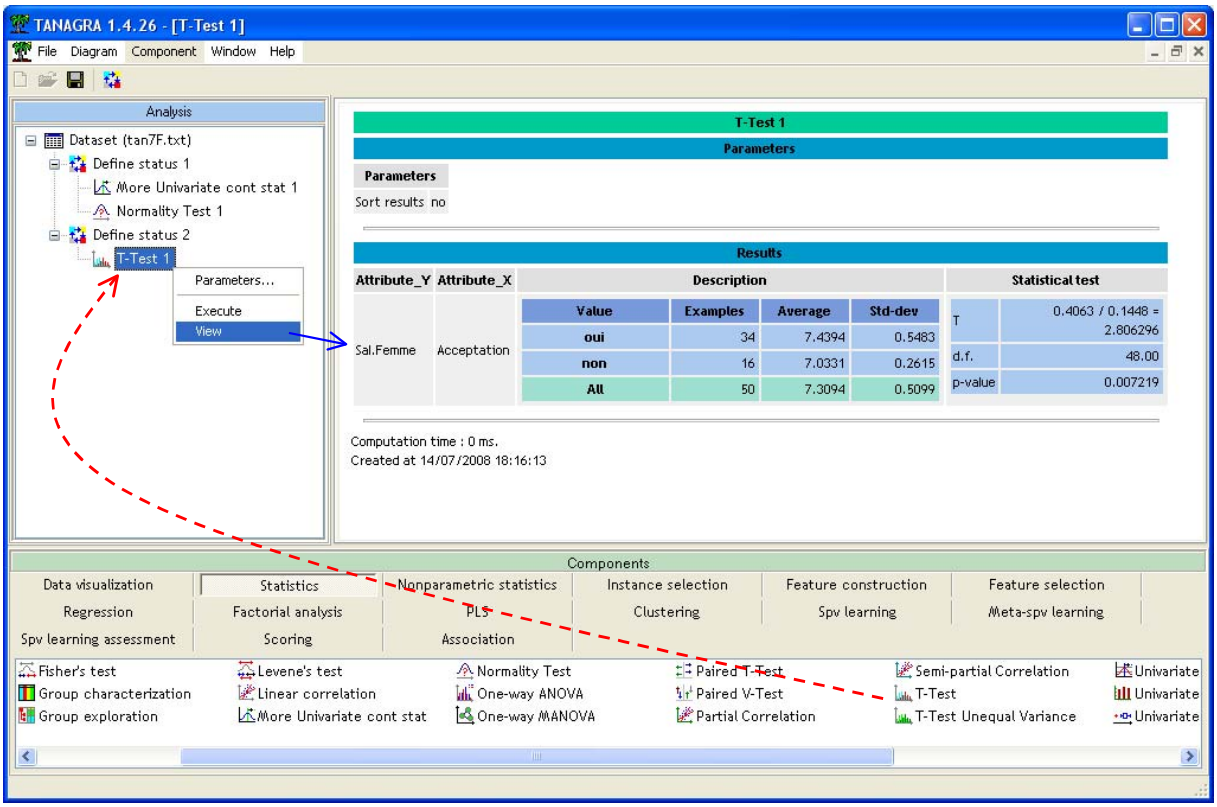
Nous souhaitons comparer le salaire féminin selon que l'acceptation. En d'autres termes, nous cherchons à savoir si l'acceptation du crédit pourrait être conditionnée par le salaire de la patronne[7]. Nous sommes dans un schéma de comparaison de moyennes, connu sous le nom de « Test de Student ».

Nous insérons un nouveau composant DEFINE STATUS dans le diagramme, nous désignons SAL.FEMME en TARGET et ACCEPTATION en INPUT.

[7] La lecture inverse est moins crédible. L'acceptation ou pas d'une demande de prêt ne devrait pas peser sur les salaires. Cet exemple illustre parfaitement le rôle d'un test statistique. La procédure calcule mécaniquement les écarts et cherche à savoir si elle est significative selon un schéma probabiliste. L'interprétation en revanche ne peut reposer que sur nos connaissances du domaine.



Puis, nous plaçons le composant T-TEST (onglet STATISTICS). Nous activons le menu VIEW.

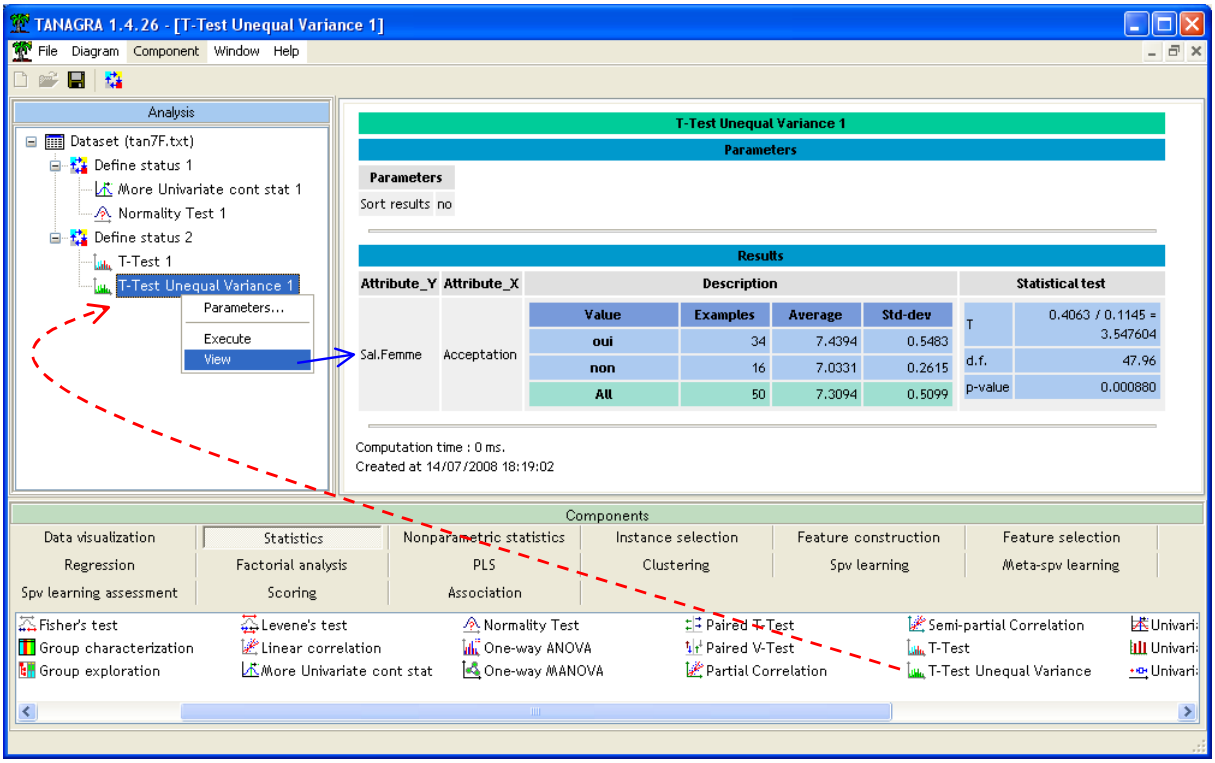


La statistique du test est t = 0.4063 / 0.1448 = 2.8063. La différence est très significative puisque la p-value est égale à 0.007219. L'acceptation du crédit est bien liée au salaire féminin.

Il faut néanmoins prendre avec précaution ce résultat. Les écarts types conditionnels sont assez différents, ils vont du simple au double. Et les effectifs sont déséquilibrés. Les conditions d'application du test de Student ne semblent pas réunies. Nous devons affiner l'analyse en introduisant la variante du test pour variances conditionnelles inégales.

4.2 Comparaison de 2 moyennes – Variances inégales

Cette variance calcule différemment l'écart type de l'écart. Elle corrige également les degrés de liberté de la loi associée. Nous insérons le composant T-TEST UNEQUAL VARIANCE (onglet STATISTICS) dans le diagramme. Nous obtenons les résultats suivants.

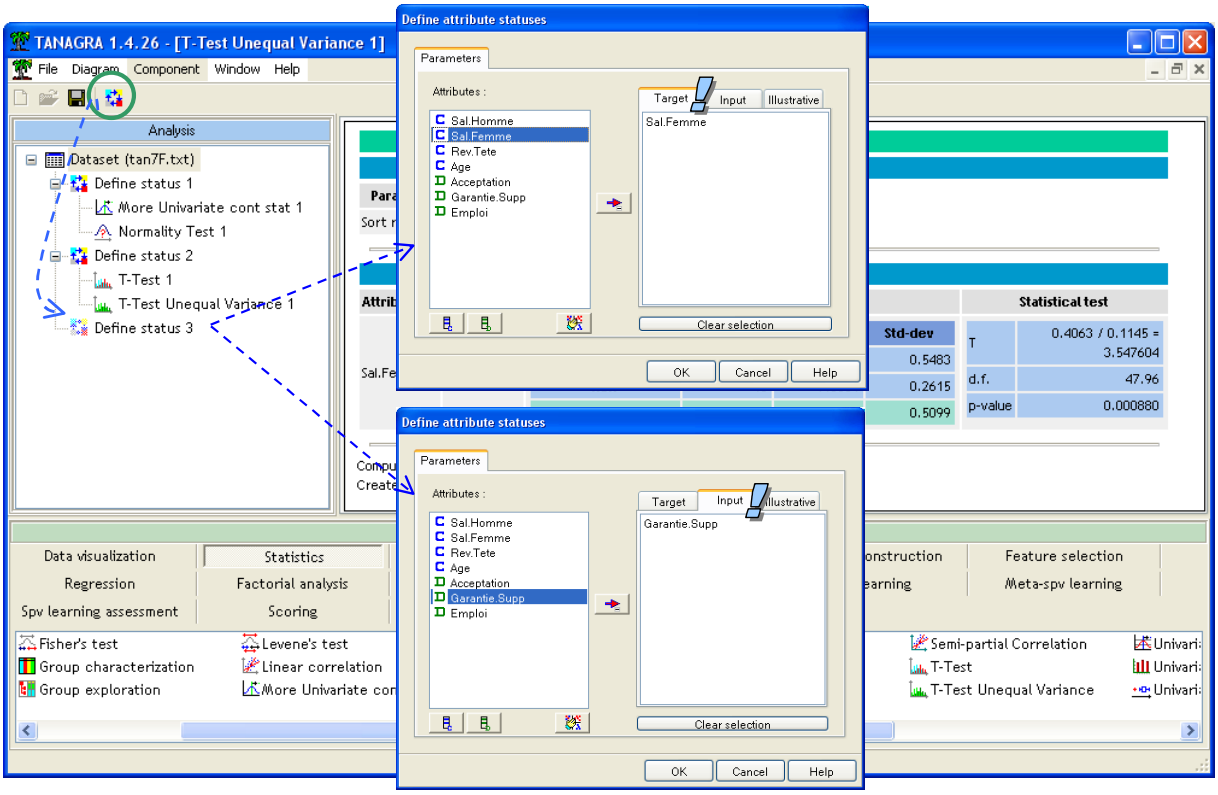


Par rapport au résultat précédent, le numérateur de la statistique du test n'est pas modifié, il en est autrement en ce qui concerne le dénominateur. Nous obtenons ainsi t = 0.4063 / 0.1145 = 3.5476. Les degrés de liberté sont d.f. = 47.96, soit en arrondissant à l'entier le plus proche d.f. ≈ 48. La p-value du test est plus faible avec p-value = 0.00088. L'écart des salaires moyens selon l'acceptation du crédit est confirmé.

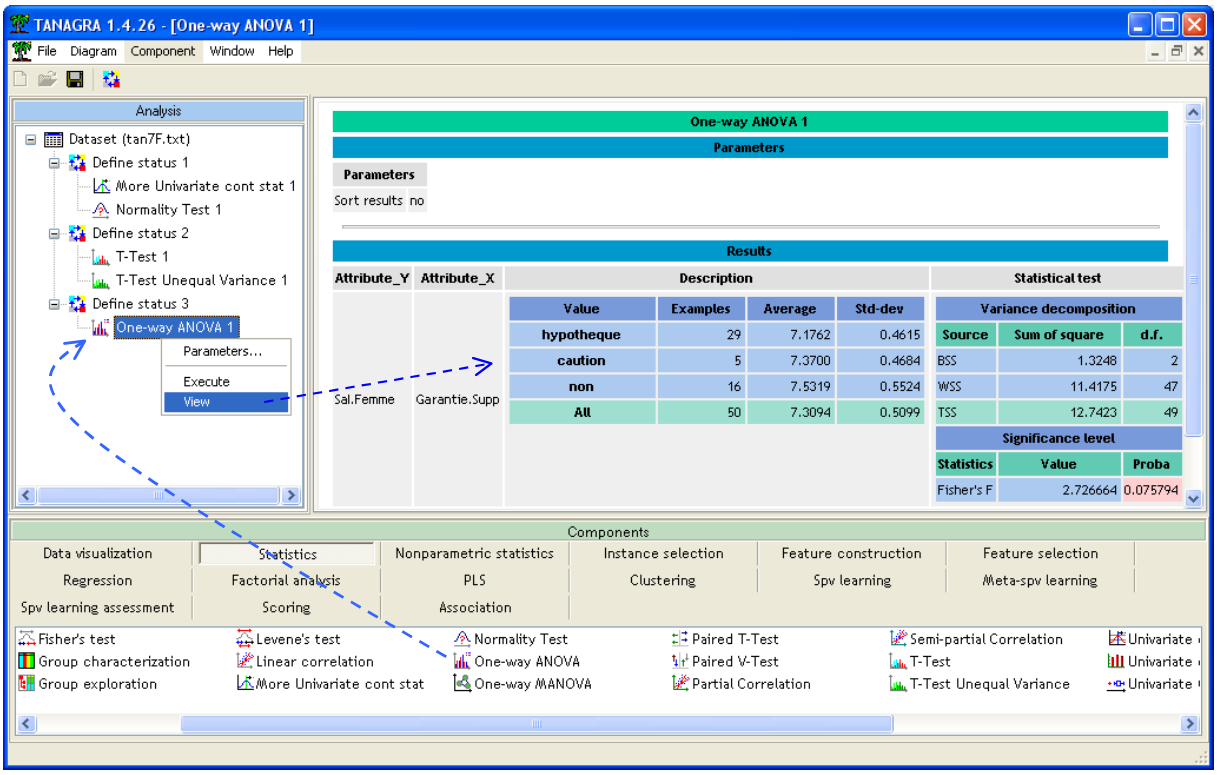
4.3 Comparaison de K moyennes – Variance égales (ANOVA)

On peut voir l'analyse de variance (ANOVA) de différentes manières. Nous la prenons sous l'angle de la généralisation du test de comparaison de moyennes dans ce didacticiel. Nous souhaitons comparer les salaires féminins dans 3 groupes définis par le type de garantie supplémentaire contractée par les emprunteurs (GARANTIE.SUPP). Dans un premier temps, nous considérons que les variances sont identiques dans les groupes.

Nous insérons le composant DEFINE STATUS dans le diagramme : SAL.FEMME est définie comme TARGET, GARANTIE.SUPP comme INPUT.



Nous introduisons alors le composant ONE-WAY ANOVA (onglet STATISTICS). Nous obtenons les résultats suivants.



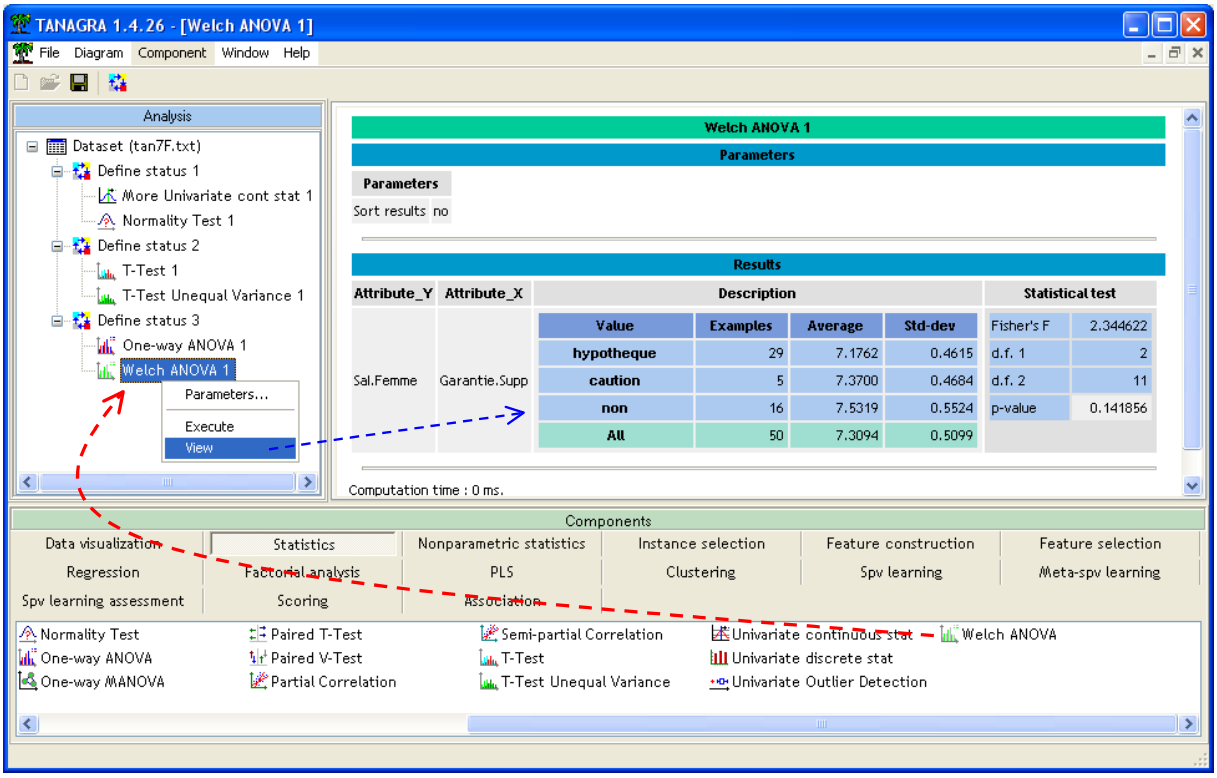
Nous observons dans le tableau « Description » les moyennes et écarts type conditionnels ; dans le la partie « Statistical Test », le tableau d'analyse de variance. La statistique F = 2.7267 ; les différences ne sont pas significatives au risque 5%, la p-value est égale à 0.07579 (elle l'est à 10%).

Même si les variances conditionnelles semblent similaires, les effectifs sont assez déséquilibrés dans les sous groupes. Les résultats de l'ANOVA sont sujets à caution dans ce cas, elle est très sensible à l'hétéroscédasticité. Voyons ce qu'il en est avec la variante de Welch.

4.4 Comparaison de K moyennes – Variances inégales – ANOVA de Welch

Cette variante est adaptée aux situations où les variances sont différentes dans les sous groupes. Voyons si les résultats précédents sont remis en cause.

Nous insérons le composant WELCH ANOVA (onglet STATISTICS) dans le diagramme.

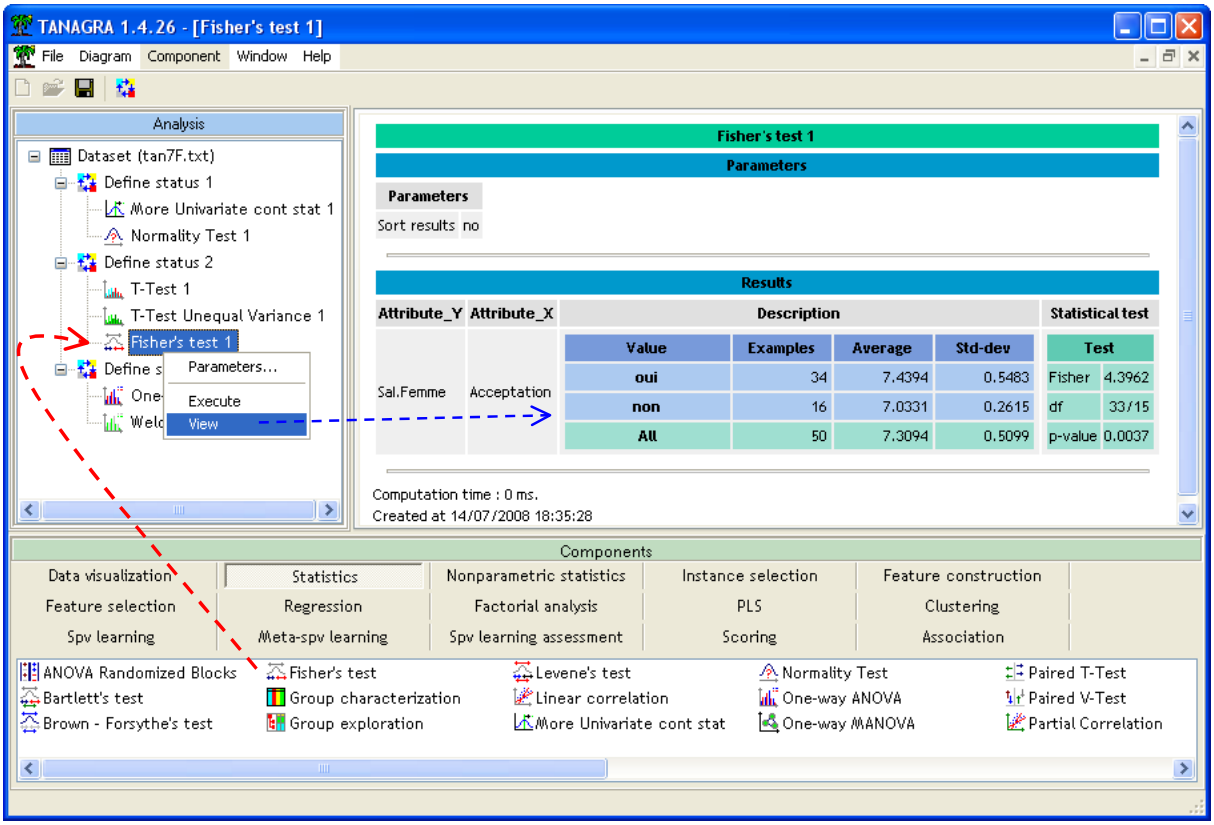


La statistique F est égale à 2.3446. Le premier degré de liberté est égal au nombre de groupes moins 1 (d.f.1 = 2) ; le second degré de liberté est plus difficile à produire, elle peut être fractionnaire, nous l'arrondissons à l'entier le plus proche (d.f.2 ≈ 11). La p-value du test est 0.141856. Les différences entre les moyennes ne sont pas significatives, même à 10%.

4.5 Comparaison de 2 variances – Test de Fisher

Les tests de comparaison de variances sont souvent utilisés préalablement aux tests de comparaison de moyennes. Ils permettraient ainsi de déterminer la procédure la plus appropriée. Mais leur rôle ne peut pas être confiné à cette tâche. Ils peuvent constituer la finalité de l'étude. Un exemple très parlant est l'étude des notes des étudiants selon leur disposition dans une salle de cours (en cercle ou en rangée). La comparaison des variabilités des notes est au moins aussi intéressante que l'étude de l'écart entre les moyennes.

Dans notre fichier de données, nous souhaitons comparer la variance du salaire féminin selon l'acceptation du crédit en utilisant le test de Fisher. Nous insérons le composant FISHER'S TEST (onglet STATISTICS) dans le diagramme, au même niveau que les composants T-TEST et T-TEST UNEQUAL VARIANCE. Ce faisant, nous bénéficions de la même spécification du rôle des variables c.-à-d. SAL.FEMME en TARGET, ACCEPTATION en INPUT.



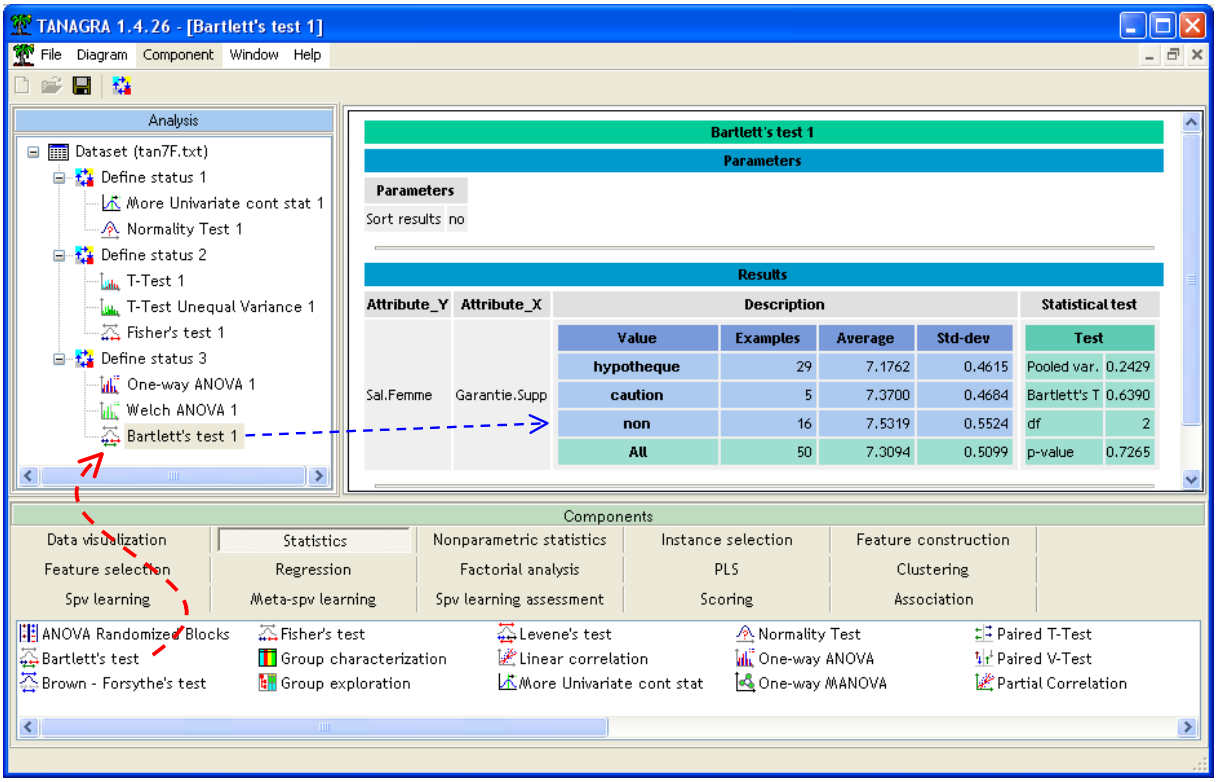
La statistique du test est F = 4.3962. Elle suit une loi de Fisher à 33 et 15 degrés de liberté. L'écart est significatif avec une p-value égale à 0.0037. Les variances sont différentes dans les sous groupes. Concernant les tests de comparaison de moyennes, la variante intégrant la différence entre les variances conditionnelles semble donc la plus appropriée, a posteriori.

4.6 Comparaison de K variances – Test de Bartlett

Ce test s'applique à la comparaison de K ($K \geq 2$) variances conditionnelles. Nous pouvons la considérer comme une généralisation du test de Fisher.

Nous souhaitons maintenant comparer la variance des salaires selon le type de garantie. Nous plaçons le composant BARTLETT'S TEST (onglet STATISTICS) dans le diagramme, à la même hauteur que les composants ONE-WAY ANOVA et WELCH ANOVA.

La statistique de Bartlett est T = 0.6390. Les écarts ne sont pas significatifs du tout, avec un p-value = 0.7265. A posteriori, il semble donc que l'ANOVA standard était la plus adaptée pour comparer les salaires selon la garantie supplémentaire apportée par les demandeurs de crédit.



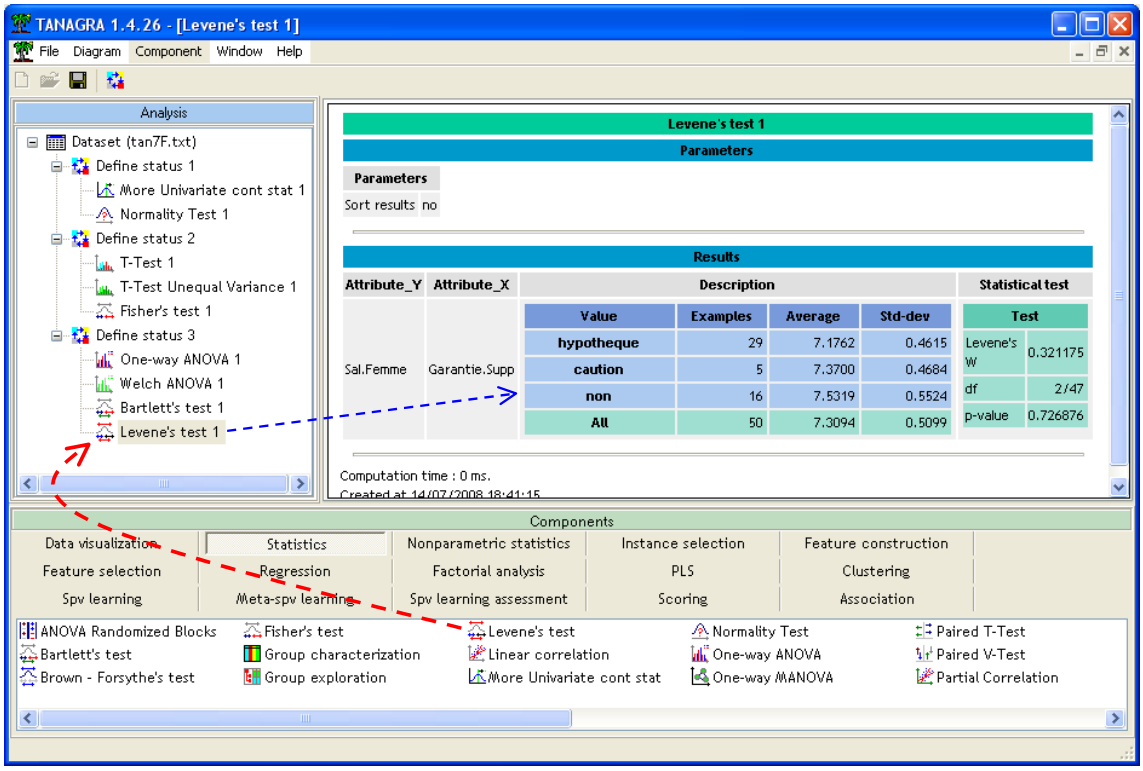
4.7 Comparaison de variances – Des tests plus robustes

Les tests de Fisher et de Bartlett sont très peu robustes lorsque les distributions s'écartent de la loi normale. Il ne devrait pas y avoir de problèmes en ce qui nous concerne. Les tests d'adéquation ont montré que les distributions des variables d'intérêt sont compatibles avec la loi normale (voir 3.3). Mais dans un cadre générique, lorsque nous avons des doutes, nous avons intérêt à utiliser des tests plus robustes, qui tiennent la route même lorsque les distributions sont asymétriques. Ils sont valables pour $K \geq 2$.

4.7.1 Le test de Levene

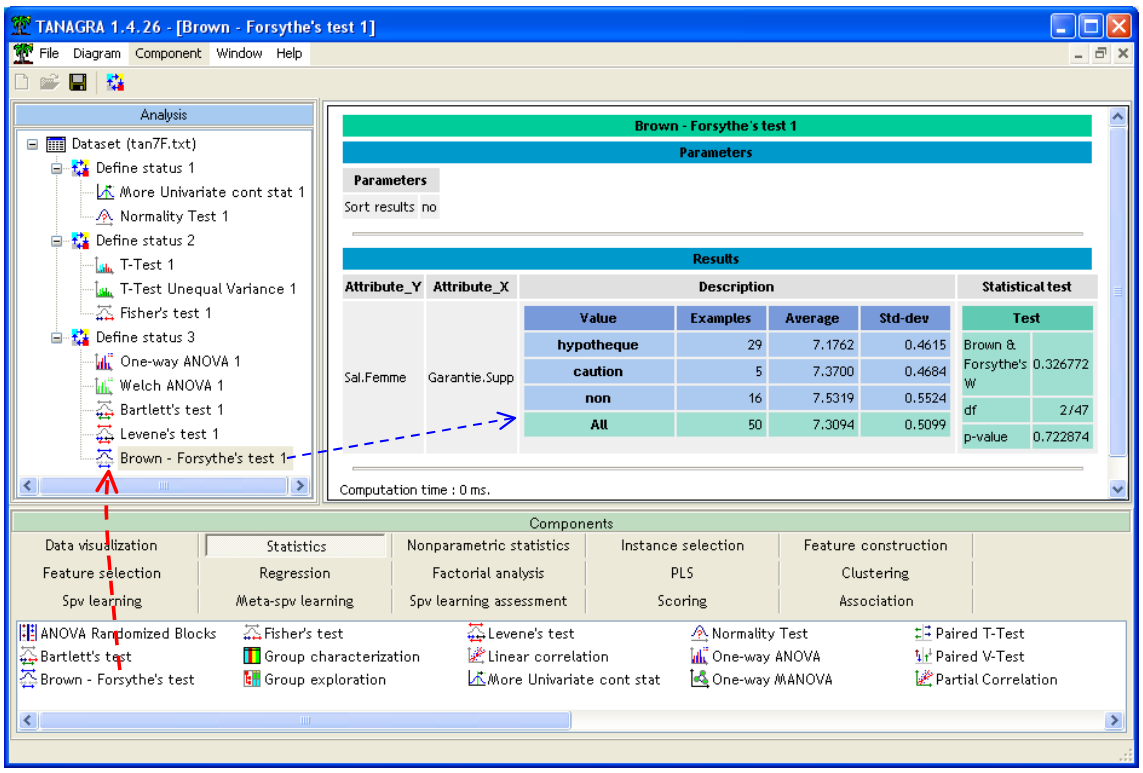
Le test de Levene est robuste face à un écart à la normalité. Il reste néanmoins performant si la distribution est gaussienne. Nous l'introduisons dans notre diagramme (composant LEVENE'S TEST, onglet STATISTICS) au même niveau que le test de Bartlett.

La statistique W est égale à 0.321175. Sous Ho, elle suit une loi de Fisher à (2 ; 47) degrés de liberté. La p-value = 0.726876, les écarts entre les variances conditionnelles ne sont pas significatifs. On notera la similitude des résultats avec le test de Barlett, sachant que la normalité des distributions semble avérée.



4.7.2 Le test de Brown-Forsythe

Cette procédure, variante du test de Levene, est certainement la plus robuste. Il faut l'utiliser en priorité si nous n'avons pas de connaissances précises sur la distribution des données. Nous plaçons le composant BROWN-FORSYTHE'S TEST (onglet STATISTICS) dans le diagramme.



Les résultats sont cohérents avec les tests de Levene et de Bartlett.

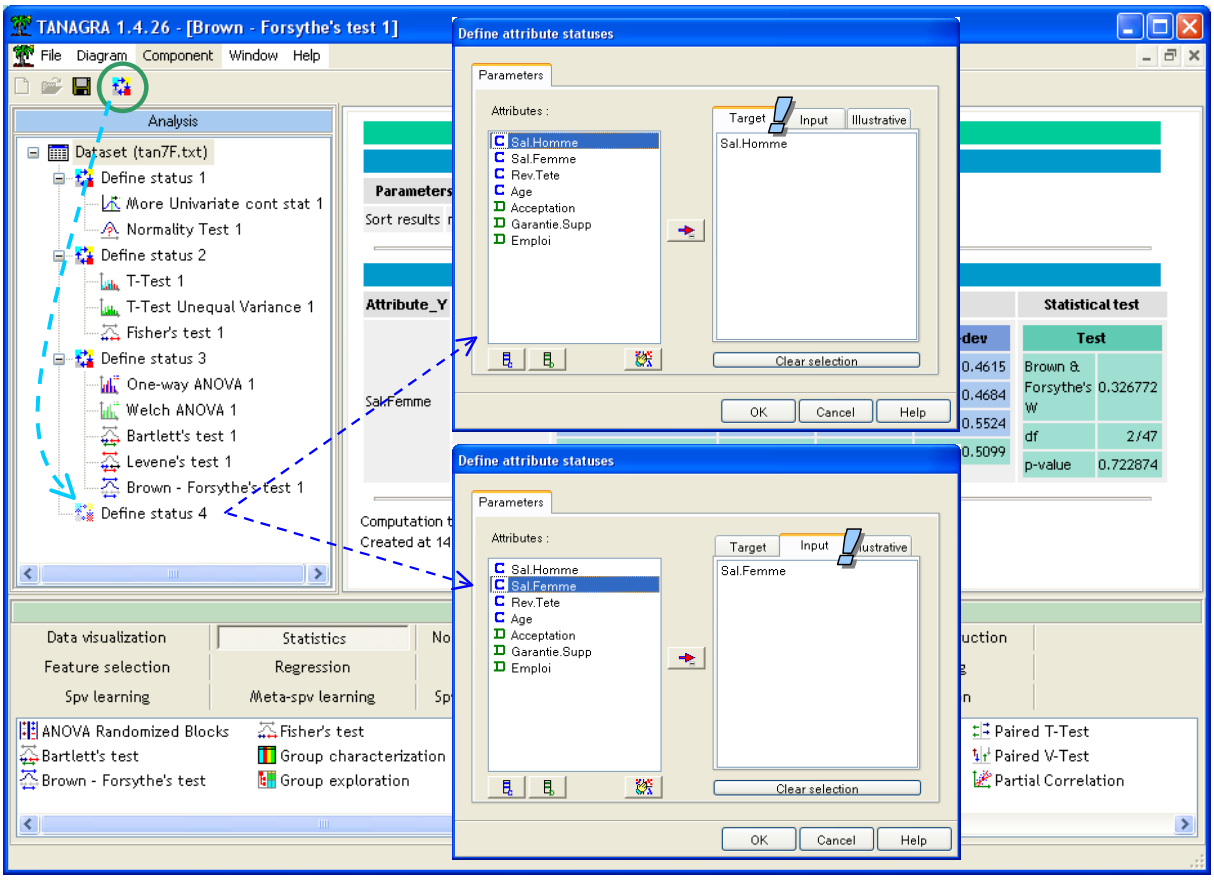
5 Tests pour échantillons appariés

L'appariement vise à réduire la variabilité (l'incertitude) due aux observations. Elle permet d'affiner les résultats[8]. On parle aussi d'échantillons liés dans la littérature.

5.1 Comparaison de moyennes pour 2 échantillons appariés

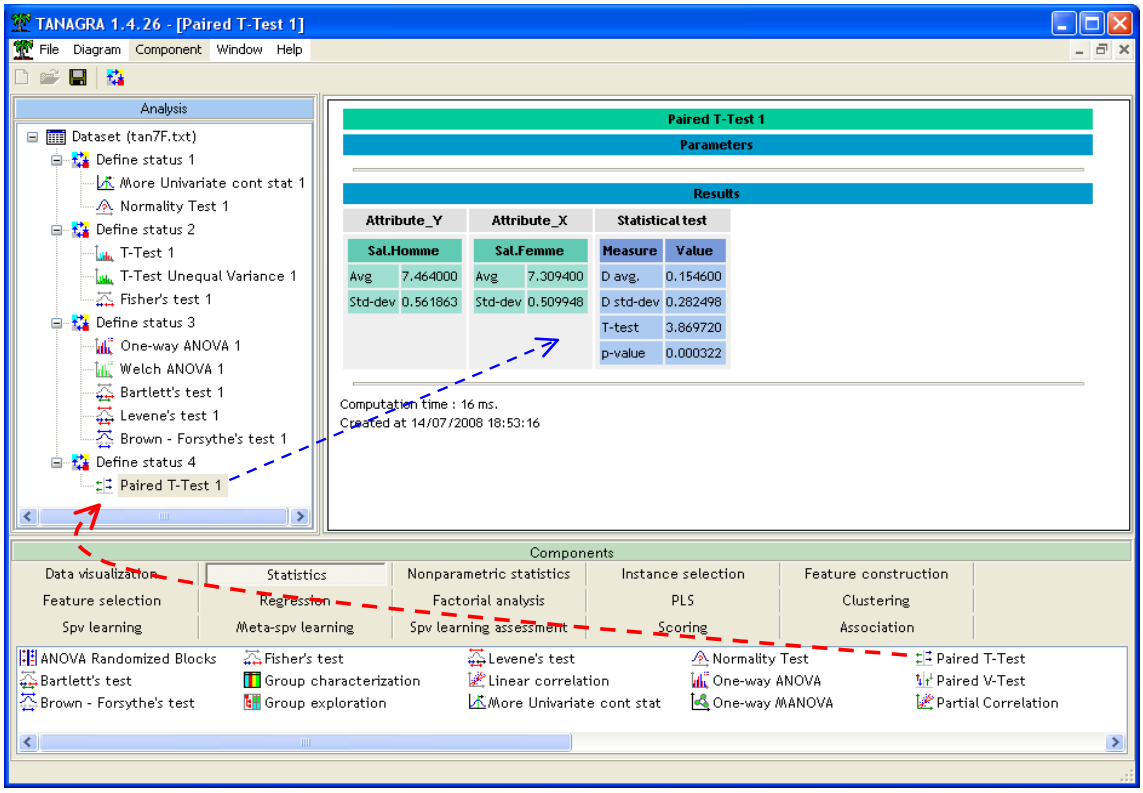
Nous souhaitons savoir si dans un ménage, l'homme a tendance à avoir un salaire plus élevé que sa femme. Il ne faut surtout pas mettre en œuvre un test pour échantillons indépendants c.-à-d. comparer la moyenne du salaire des hommes avec la moyenne de celui des femmes. En effet, la confrontation doit se faire à **l'intérieur des ménages**. Nous sommes dans un schéma de test pour échantillons appariés.

Nous déposons un nouveau DEFINE STATUS dans le diagramme, nous plaçons en TARGET la variable SAL.HOMME et en INPUT SAL.FEMME.



Nous introduisons alors le composant PAIRED T-TEST (onglet STATISTICS).

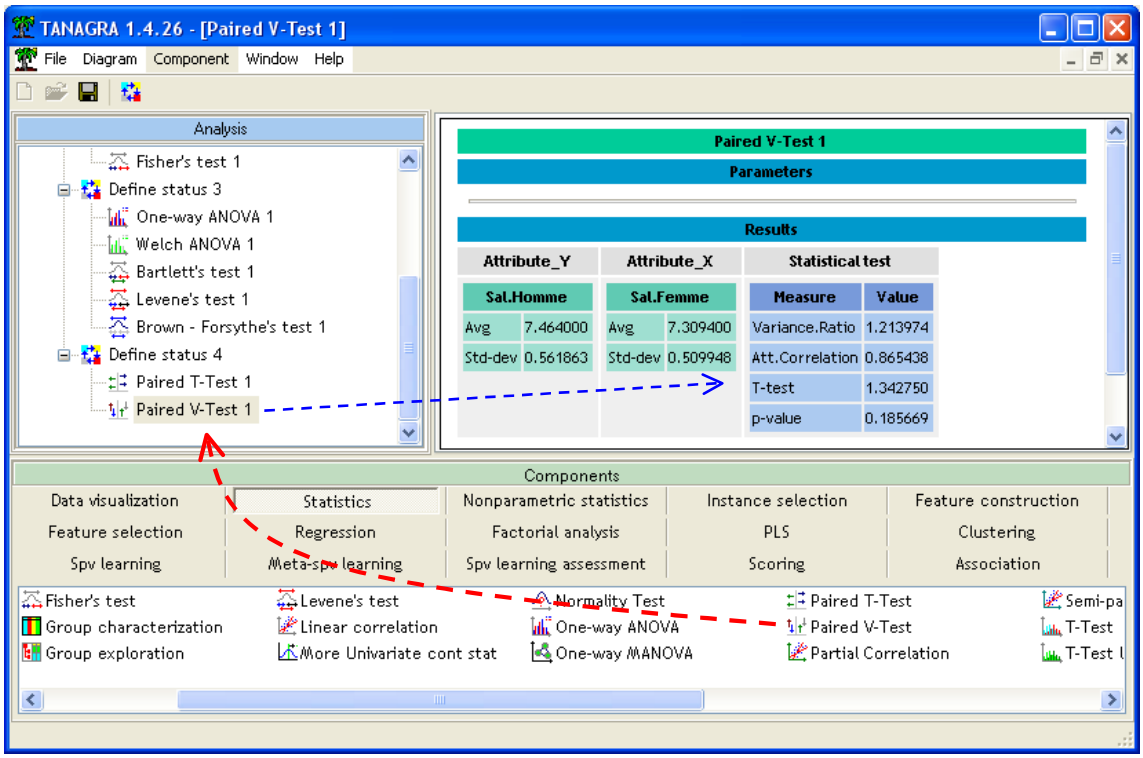
[8] Pour plus de précisions sur la démarche, voir notre support de référence http://eric.univ-lyon2.fr/~ricco/cours/cours/Comp_Pop_Tests_Parametriques.pdf ; Partie 2.



La moyenne de l'écart entre salaire masculin et féminin dans les ménages est de D.AVG = 0.1546. La statistique du test est t = 3.86972. La différence est significative avec une p-value de 0.000322.

5.2 Comparaison de variances pour 2 échantillons appariés

Nous pouvons appliquer la même démarche pour la comparaison des variances. Au même niveau que le composant précédent, nous insérons PAIRED V-TEST (onglet STATISTICS).



L'écart type des salaires masculins est 0.561863, il est de 0.509948 pour les femmes. Le ratio des variances est donc 0.561863^2 / 0.509948 = 1.213974. Mais les salaires sont fortement corrélés avec r = 0.865438. La statistique du test est alors égal à t = 1.342750. Au niveau de signification 5%, les salaires sont homogènes, la p-value du test est égale à 0.185669.

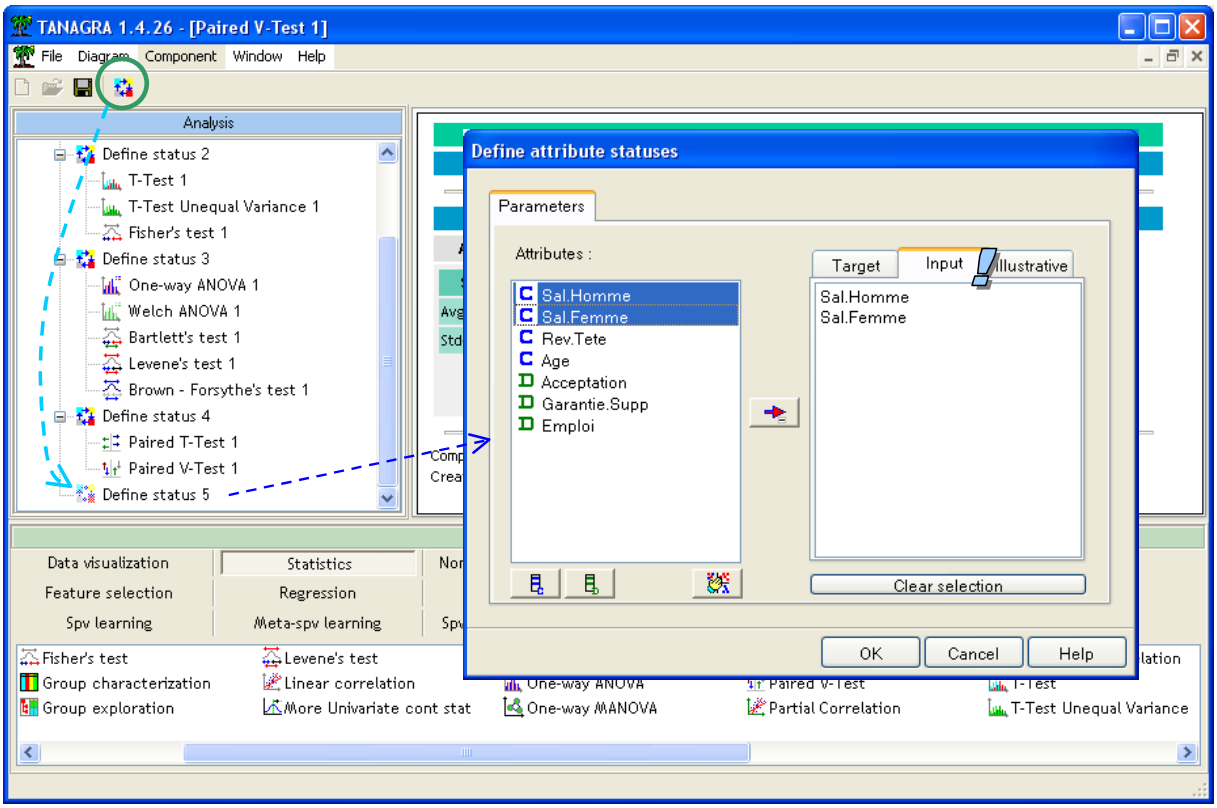
5.3 Comparaison de moyennes pour K échantillons appariés

Nous pouvons étendre la comparaison de moyennes pour K échantillons appariés. Dans les schémas d'expérimentation, on parle de « blocs aléatoires complets ». Il s'agit simplement de généraliser le cadre précédent à K > 2 échantillons.

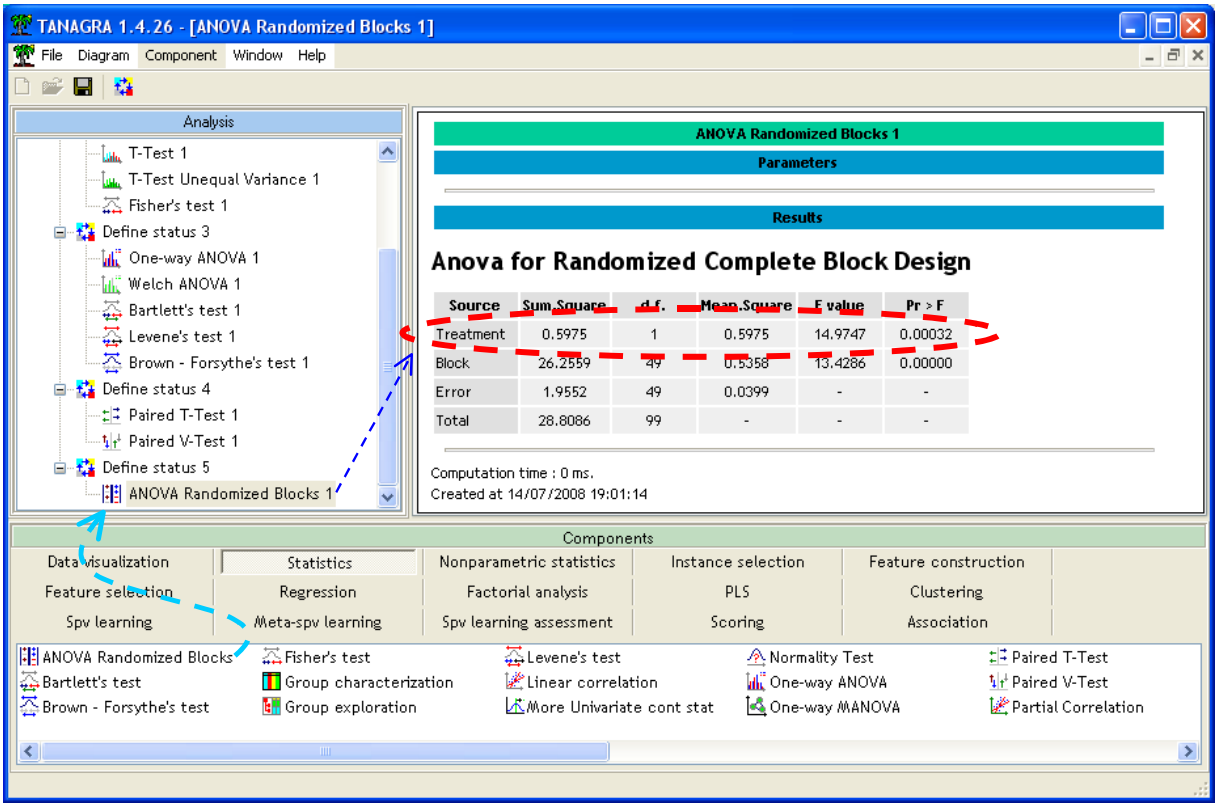
Nous allons mettre en œuvre le composant ANOVA RANDOMIZED BLOCKS (onglet STATISTICS) sur le même exemple de comparaison des salaires (section 5.1). Nous pourrons ainsi faire le parallèle entre les 2 approches.

La sélection de variables est un peu différente avec ce composant. Nous insérons DEFINE STATUS dans le diagramme, nous plaçons en INPUT les variables SAL.HOMME et SAL.FEMME.

Nous pouvons placer autant de variables que l'on veut en INPUT. Il faut simplement qu'elles soient de type continu (C).



Puis nous introduisons ANOVA RANDOMIZED BLOCKS.



Ce qui nous intéresse au premier chef est la source de variabilité expliquée par les « traitements » c.-à-d. la différence entre SAL.HOMME et SAL.FEMME. La statistique du test F = 14.9747 suit une loi de Fisher à (1 ; 49) degrés de liberté. Ici également, l'écart s'avère significatif avec une p-value de 0.00032.

Equivalence des tests de comparaison de moyennes pour échantillons appariés lorsque K = 2. Nous savons qu'il y a une équivalence entre la loi de Student et la loi de Fisher. En effet, $[T(m)]^2 \equiv F(1,m)$. Concernant notre statistique de test, nous trouvons ainsi $\sqrt{F} = \sqrt{17.9747} = 3.86972 = t$, soit la valeur de la statistique du test de comparaison de 2 moyennes pour échantillons appariés (section 5.1).