Objectif

Statistique descriptive avec TANAGRA.

La statistique descriptive vise à résumer l'information portée par un tableau de données. « Trop d'informations tue l'information » a-t-on coutume de dire. Il est illusoire d'inspecter un tableau contenant des centaines, voire des milliers, d'observations et d'en déduire des tendances.

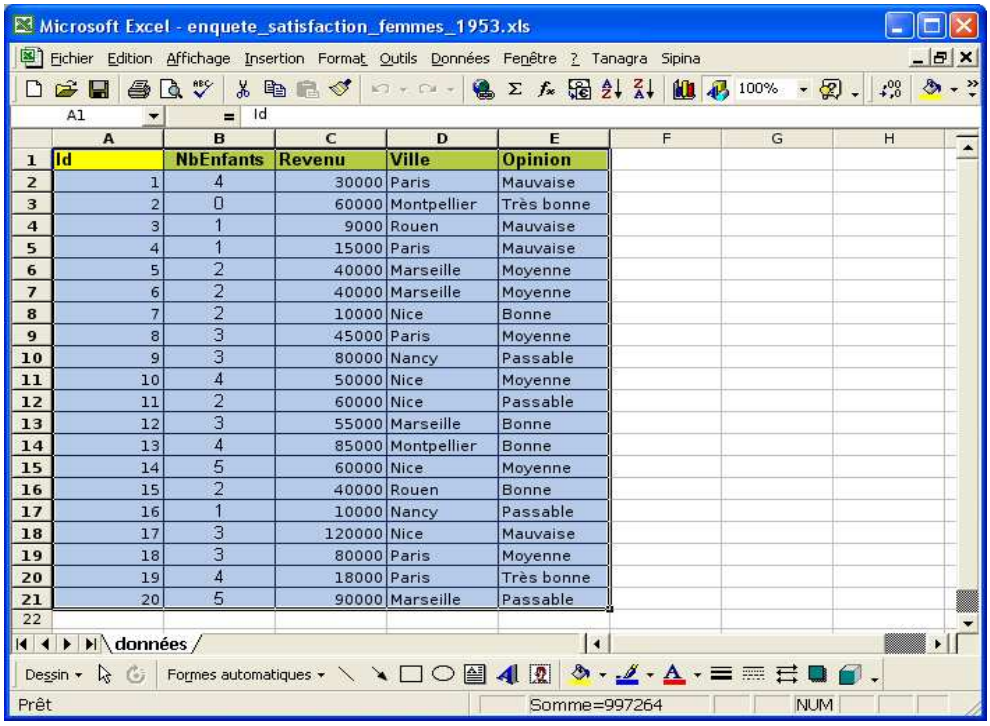
L'objectif de la statistique descriptive est de nous fournir une image simplifiée de la réalité, en mettant en exergue des caractéristiques qui ne sont pas discernables de prime abord. Elle emmène un nouvel éclairage sur les données. Elle s'appuie pour cela sur des indicateurs et des représentations graphiques qui, pour simples qu'elles soient, sont très souvent pertinentes pour une bonne compréhension de la structure des données.

Nous présenterons les techniques descriptives selon deux axes. Tout d'abord nous ferons la distinction « techniques univariées », qui étudient les variables individuellement, et « techniques bivariées », qui étudient les relations entre 2 variables. Le second axe repose sur la distinction entre les variables catégorielles (qualitatives nominales) et les variables continues (quantitatives).

Données

Fichier de données

Les données décrivent les résultats d'une enquête de satisfaction par rapport à un produit pour bébé quelconque auprès de 20 femmes. Elles proviennent du site de Fabrice Mazerolle[1] qui, outre les données, propose un cours en ligne d'excellente facture qui peut servir de repère dans ce didacticiel. Par rapport au fichier originel, nous avons supprimé les informations nominatives (c'est le cas généralement dans le traitement des enquêtes).

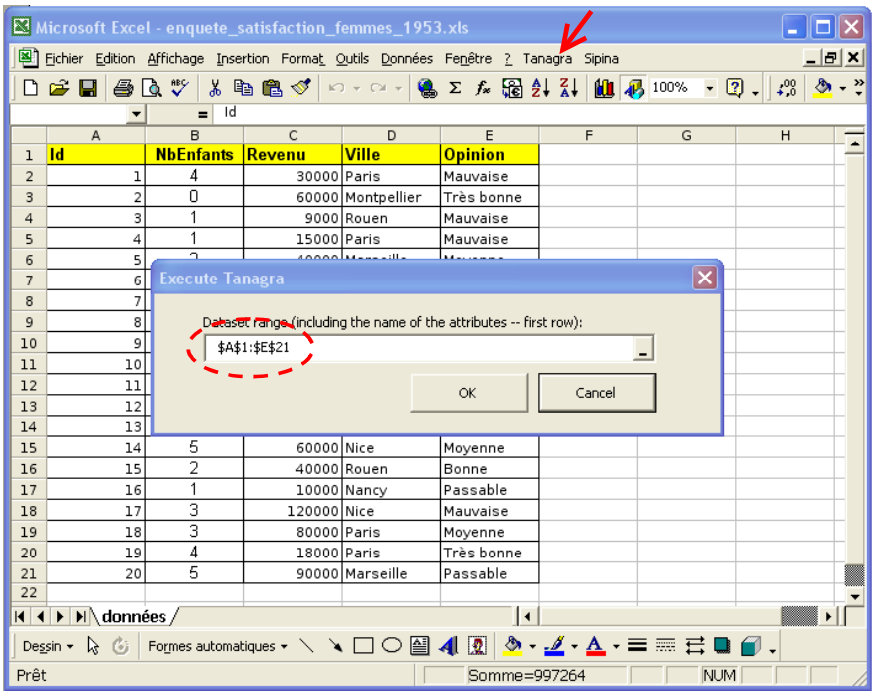


	A	B	C	D	E
1	Id	NbEnfants	Revenu	Ville	Opinion
2	1	4	30000	Paris	Mauvaise
3	2	0	60000	Montpellier	Très bonne
4	3	1	9000	Rouen	Mauvaise
5	4	1	15000	Paris	Mauvaise
6	5	2	40000	Marseille	Moyenne
7	6	2	40000	Marseille	Moyenne
8	7	2	10000	Nice	Bonne
9	8	3	45000	Paris	Moyenne
10	9	3	80000	Nancy	Passable
11	10	4	50000	Nice	Moyenne
12	11	2	60000	Nice	Passable
13	12	3	55000	Marseille	Bonne
14	13	4	85000	Montpellier	Bonne
15	14	5	60000	Nice	Moyenne
16	15	2	40000	Rouen	Bonne
17	16	1	10000	Nancy	Passable
18	17	3	120000	Nice	Mauvaise
19	18	3	80000	Paris	Moyenne
20	19	4	18000	Paris	Très bonne
21	20	5	90000	Marseille	Passable
22					

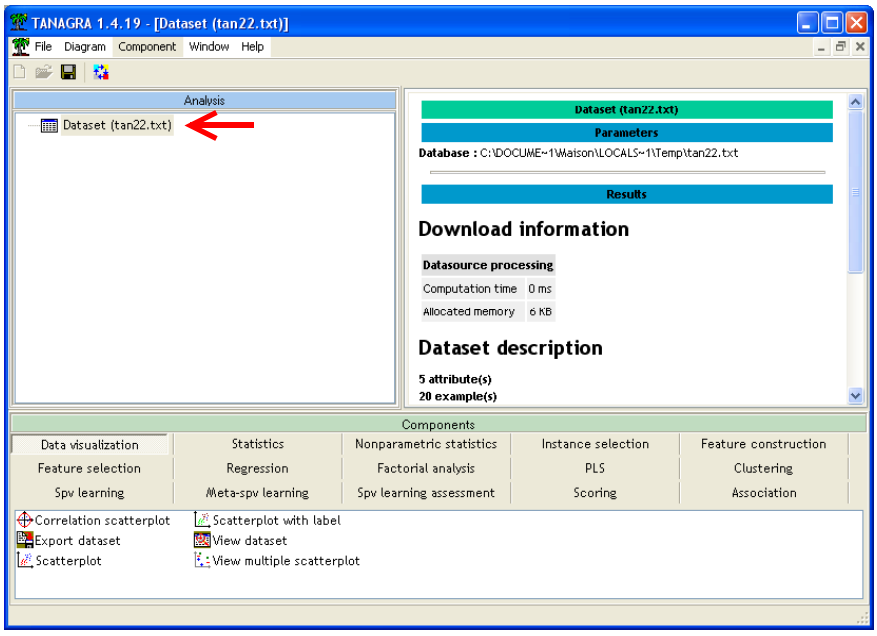
[1] http://www.mazerolle.fr/stats/stats.htm

Créer un diagramme dans TANAGRA

Dans un premier temps, il faut initialiser un diagramme de traitements et charger les données dans le logiciel TANAGRA. Le plus simple est d'ouvrir le fichier « enquete_satisfaction_femmes_1953.xls » dans le tableur EXCEL. Nous sélectionnons la plage de données et nous activons le menu TANAGRA/EXECUTE TANAGRA installée à l'aide de la macro complémentaire TANAGRA.XLA livrée avec le logiciel[2].



TANAGRA est automatiquement lancé, un nouveau diagramme de traitements est mis en place et les données sont disponibles à la racine du diagramme.



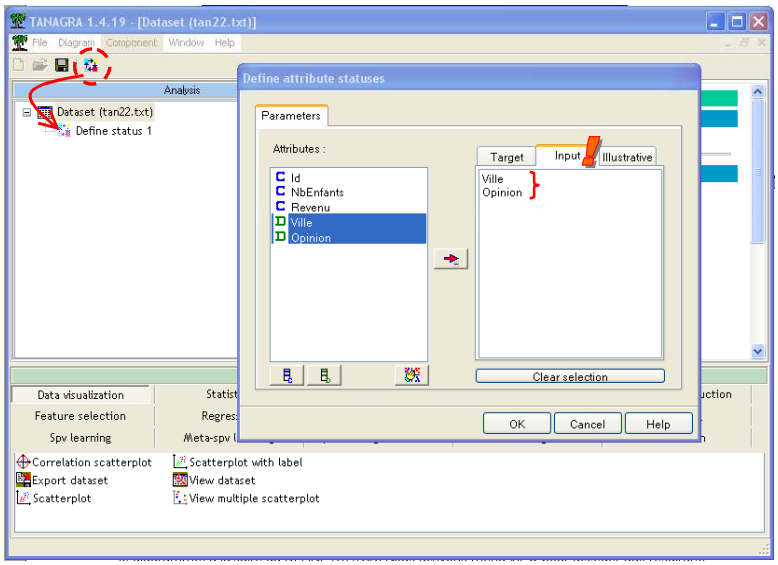
[2] Cette macro est disponible depuis la version 1.4.11 de TANAGRA. Un didacticiel décrit la procédure d'installation (http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Excel_AddIn.pdf).

Variables qualitatives

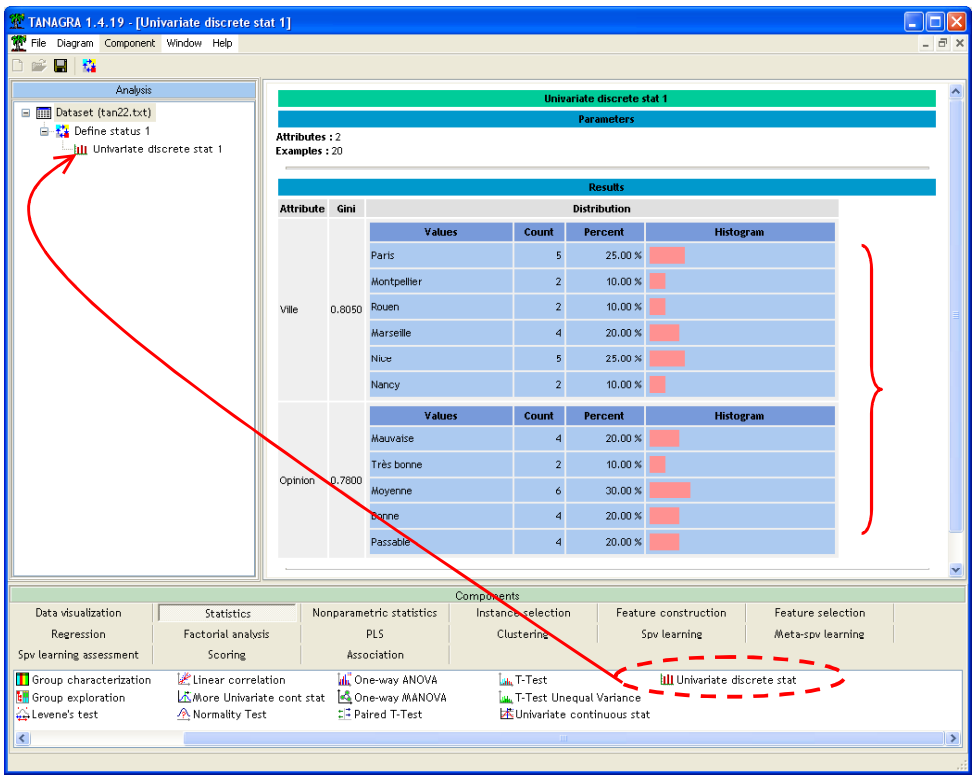
Statistiques univariées

L'outil privilégié pour décrire une variable qualitative est le diagramme en bâtons. Il s'agit d'énumérer les valeurs (les modalités) prises par la variable, et pour chaque modalité fournir l'effectif correspondant, et éventuellement la proportion (effectif modalité / effectif total) associée.

Dans TANAGRA, nous devons définir les variables à manipuler. Pour ce faire, nous introduisons le composant DEFINE STATUS en utilisant le raccourci dans la barre d'outils. La boîte de paramétrage apparaît automatiquement, nous sélectionnons en INPUT les variables VILLE et OPINION.



Nous introduisons maintenant le composant UNIVARIATE DISCRETE STAT (onglet STATISTICS) dans le diagramme, à la suite du DEFINE STATUS. Nous activons menu VIEW pour accéder aux résultats.



Nous obtenons les distributions de fréquences pour nos deux variables. Nous observons les fréquences absolues (effectif par modalité), et les fréquences relatives (pourcentages). Nous constatons ainsi, s'agissant de la variable VILLE, que 25% des femmes enquêtées viennent de Paris, 10% de Montpellier, etc.

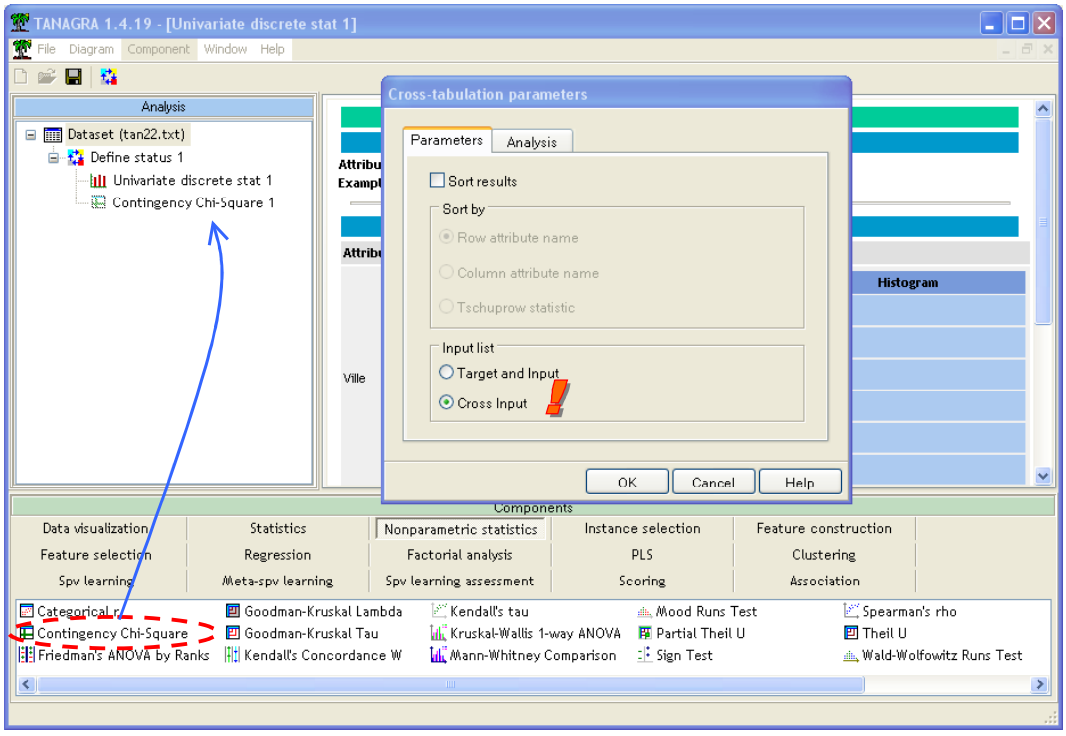
L'indice de GINI permet de situer la concentration des effectifs sur certaines modalités. Une valeur proche de 0 indique que les effectifs sont concentrés sur une des modalités, s'il tend vers (1 – 1/K) où K est le nombre de modalités, nous avons équirépartition c.-à-d. les pourcentages sont les mêmes sur l'ensemble des modalités. Dans le cas de VILLE, la valeur maximale est 1-1/6 = 0.83, la valeur observée est 0.8, cela confirme que les différentes villes sont assez équitablement représentées dans ce fichier.

Statistiques bivariées – Croisement de variables qualitatives

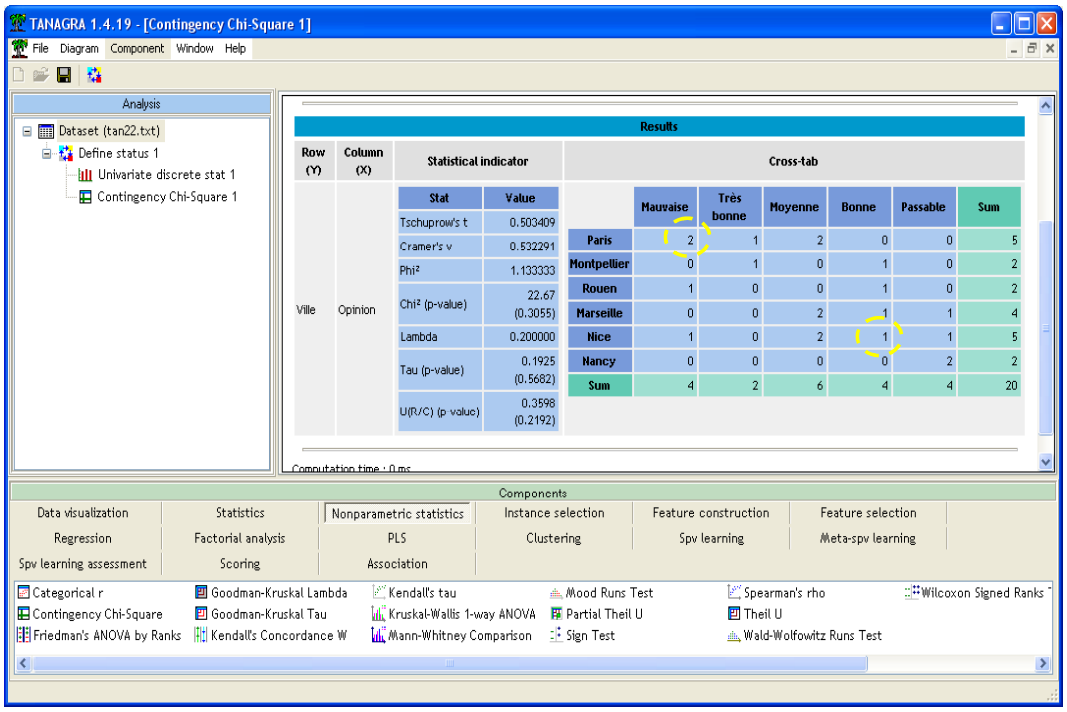
Nous nous intéressons maintenant à la répartition de l'opinion selon les villes. Nous devons former un tableau croisé, dit de contingence, il doit calculer les effectifs pour chaque association des modalités des variables.

Fréquences absolues

A la suite du DEFINE STATUS, nous plaçons le composant CONTINGENCY CHI-SQUARE (onglet NONPARAMETRIC STATISTICS). Nous activons le menu PARAMETERS, et nous indiquons que les variables sont énumérées en tant qu'INPUT (Option INPUT LIST). Dans ce cas, un tableau de contingence est élaboré pour chaque couple de variable INPUT.



Nous cliquons sur le menu VIEW et nous obtenons :



Le tableau permet la lecture suivante : 2 personnes de « Paris » ont une opinion « mauvaise » du produit, 1 personne de « Nice » a une « bonne » opinion du produit, etc.

Une pléiade d'indicateurs est affichée. *Grosso modo*, ils indiquent le lien qui existe entre les deux variables du tableau. Pour se donner une idée simple de ce qu'ils véhiculent, prenons l'exemple du T de Tschuprow. Il prend la valeur 0 si les deux variables sont indépendantes c.-à-d. la connaissance de la valeur prise par l'une ne donne aucune indication sur la valeur prise par la seconde variable (en inversement). Si la relation est déterministe, sa valeur est 1. Il existe une littérature abondante sur ce thème[3].

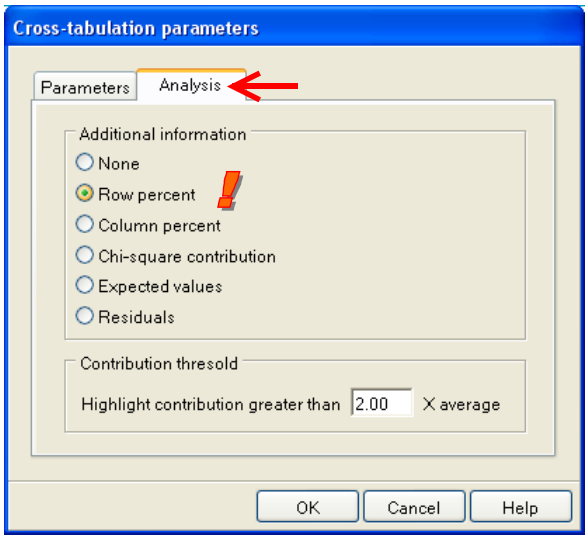
Profils lignes et colonnes

La lecture directe des valeurs dans les cases du tableau de contingence apporte peu d'informations en définitive. Il est plus intéressant de s'intéresser aux pourcentages en ligne ou en colonne.

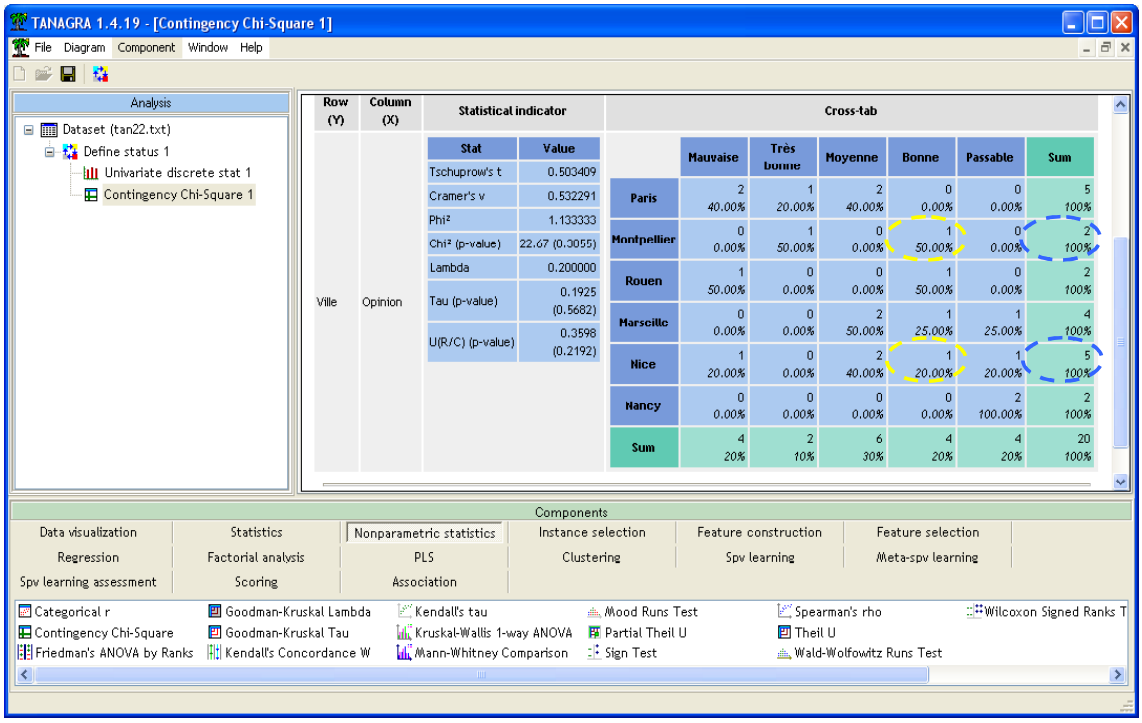
Si l'on reprend la satisfaction selon les villes, il paraît plus judicieux de ramener les données sur un même référentiel en divisant les effectifs par le nombre de personnes associées à chaque ville. Ainsi, les chiffres seront comparables. Prenons un exemple très simple : 1 personne a une bonne opinion du produit à Montpellier, il en est de même à Nice. Apparemment, la situation est identique.

Dans TANAGRA, nous pouvons afficher les profils en modifiant le paramétrage du composant. Nous cliquons sur le menu PARAMETERS. Dans l'onglet ANALYSIS, nous sélectionnons l'option ROW PERCENT. Nous validons en cliquant sur OK.

[3] http://www2.chass.ncsu.edu/garson/PA765/assocnominal.htm



En actionnant le menu VIEW, nous obtenons :



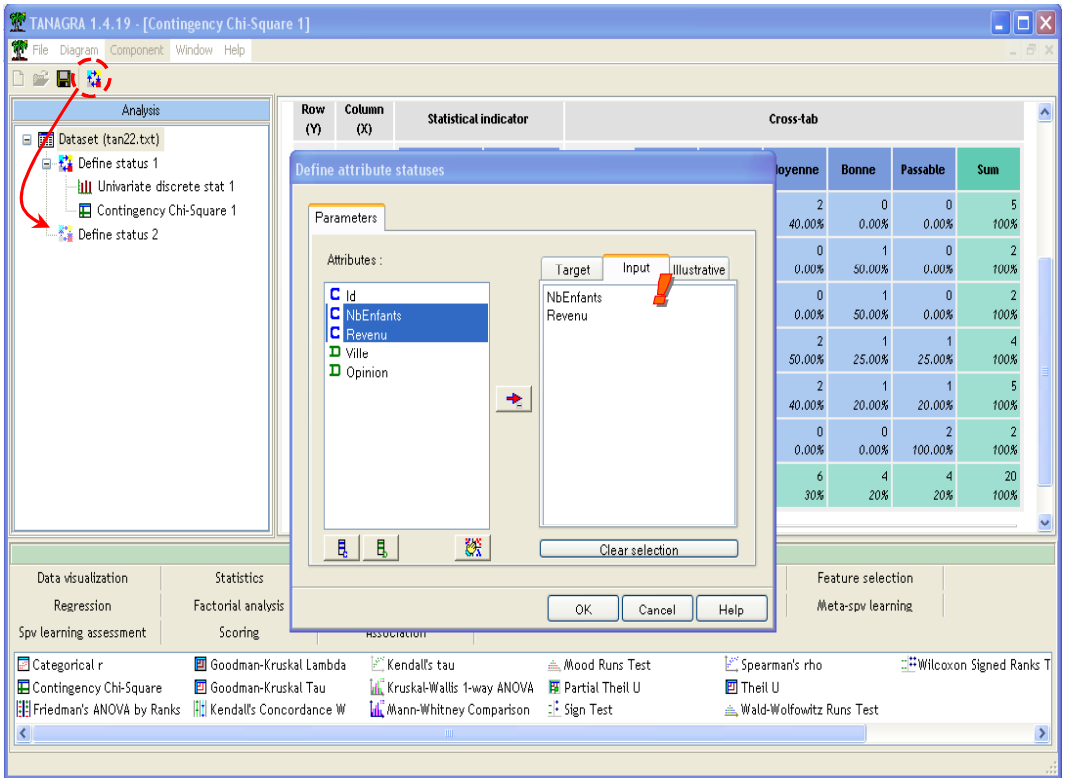
Reprenons notre exemple ci-dessus. Nous constatons que 50% des personnes ont une bonne opinion du produit à Montpellier, en revanche, ils ne sont que 20% à Nice. La raison est que l'effectif, on parle d'effectif marginal, n'est pas le même dans les deux villes.

Variables quantitatives

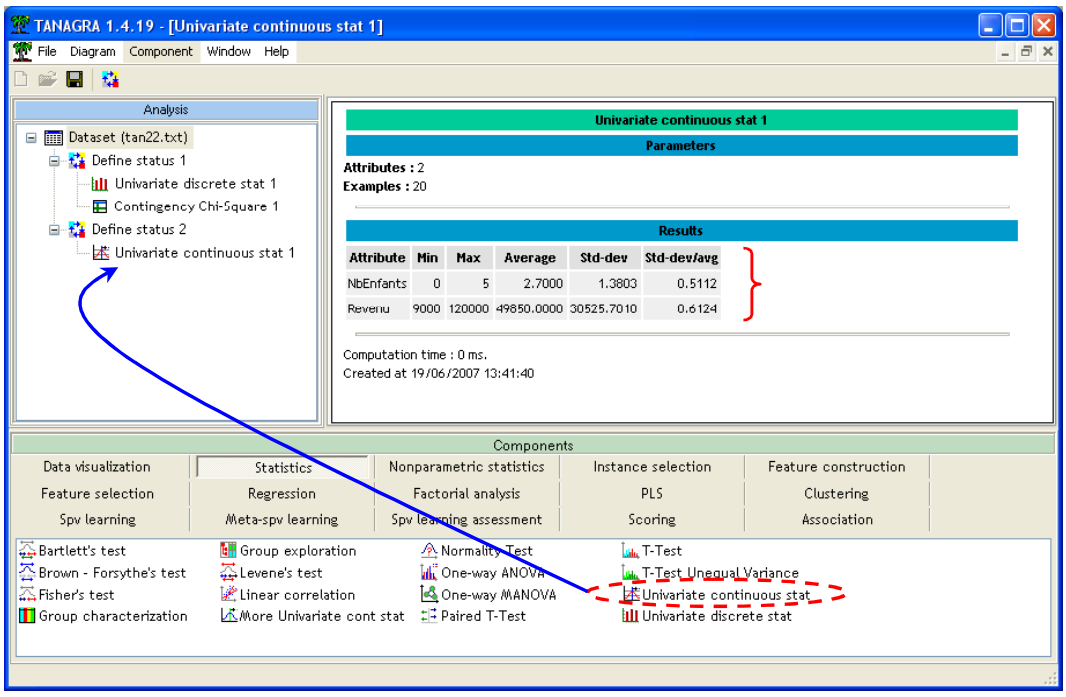
Statistiques univariées

Dans TANAGRA, les outils sont plus riches concernant les variables quantitatives. Nous disposons de deux composants qui répondent à des spécifications différentes.

Plaçons le composant DEFINE STATUS à la racine du diagramme, toujours en utilisant le raccourci de la barre d'outils, ce qui est le plus simple. Nous plaçons en INPUT maintenant les variables NBENFANTS et REVENU (calculer des statistiques sur les identifiants n'est pas très pertinent).

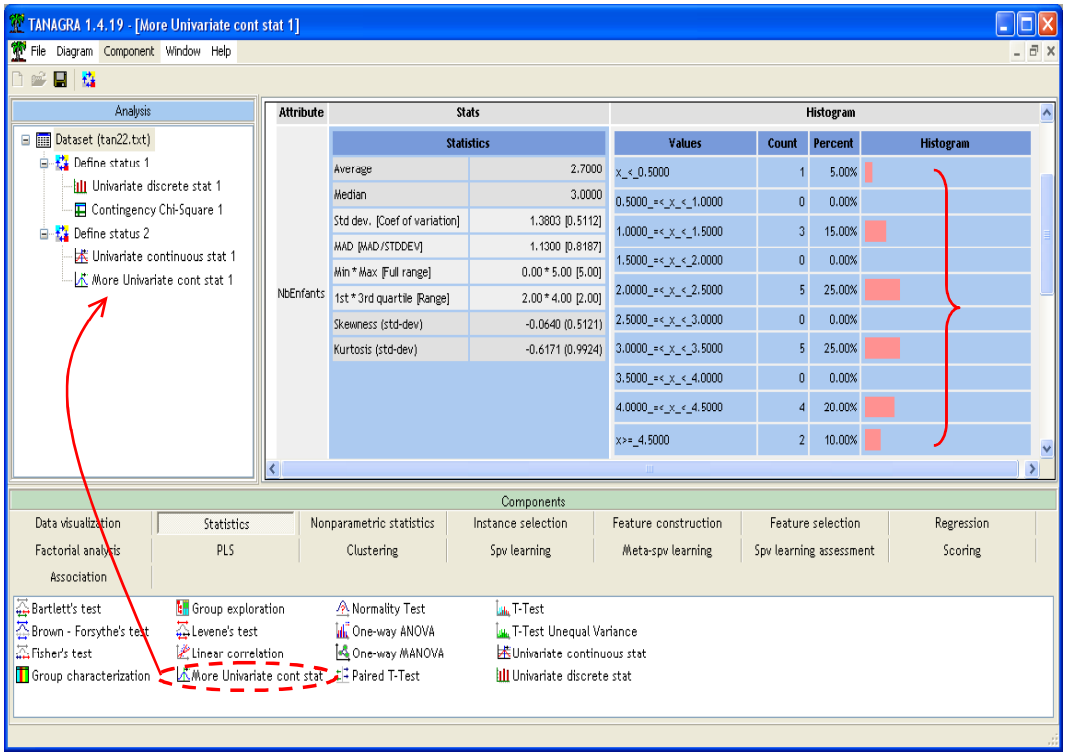


Puis nous insérons le composant UNIVARIATE CONTNUOUS STAT (onglet STATISTICS). Nous cliquons sur le menu VIEW.



Ce composant propose quelques indicateurs simples telles que la moyenne, le min, l'écart type. Il est adapté lorsque nous voulons calculer ces indicateurs sur un très grand nombre de variables et donne une idée très simplifiée de la structure des données. Par exemple, si l'écart type est égal à zéro ou, c'est la même chose, si le min est égal au max, nous pouvons affirmer que la variable n'est composée que d'une seule valeur, c'est une constante.

Un second outil fournit des informations détaillées. Nous insérons le composant MORE UNIVARIATE CONTNUOUS STAT (onglet STATISTICS) à la suite du DEFINE STATUS 2. Nous visualisons les résultats.



En plus des indicateurs précédents, de nouveaux indicateurs sur la forme de la distribution sont proposés : les coefficients d'asymétrie et d'aplatissement, etc. Un histogramme de fréquence permet de préciser la répartition des données.

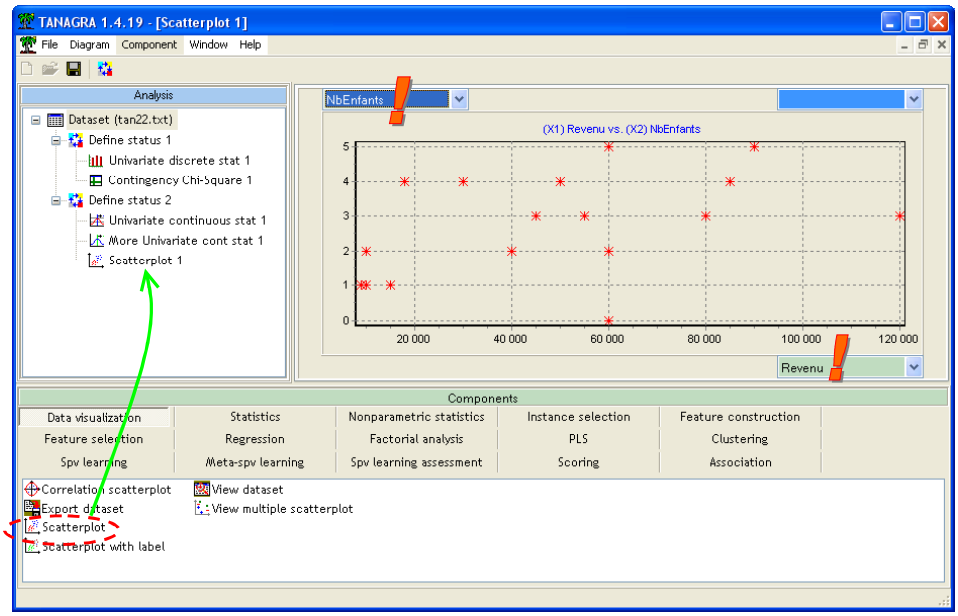
La largeur et le nombre des intervalles sont automatiquement définis par TANAGRA, il n'est pas possible de les modifier pour afficher des histogrammes avec des amplitudes inégales entre autres.

Statistiques bivariées – Croisement de variables quantitatives

Graphique nuage de points

A la différence des variables qualitatives, il existe des outils graphiques pour la visualisation des couples de variables quantitatives dans TANAGRA. Nous voulons par exemple le lien qui pourrait exister entre le nombre d'enfants et le revenu des personnes.

Nous plaçons en dessous du DEFINE STATUS 2 le composant SCATTERPLOT (onglet VISUALIZATION). En cliquant sur VIEW, le nuage de point apparaît, il est possible de modifier à la volée les variables en abscisse et en ordonnée. Pour une bonne visualisation, la taille des points dans le graphique est paramétrable.

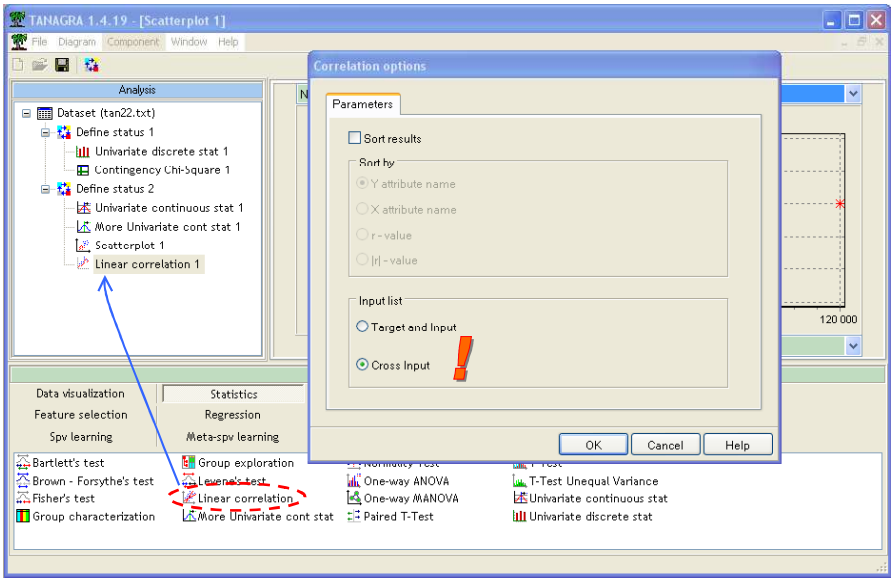


Lorsque le REVENU augmente, les femmes ont tendance à avoir un plus grand nombre d'enfants. A moins que ce soit l'inverse. C'est bien là d'ailleurs les limites de la statistique. Nous pouvons observer des régularités. Après, les transformer en causalité repose en grande partie sur l'expert du domaine, le sociologue, l'économiste, etc.

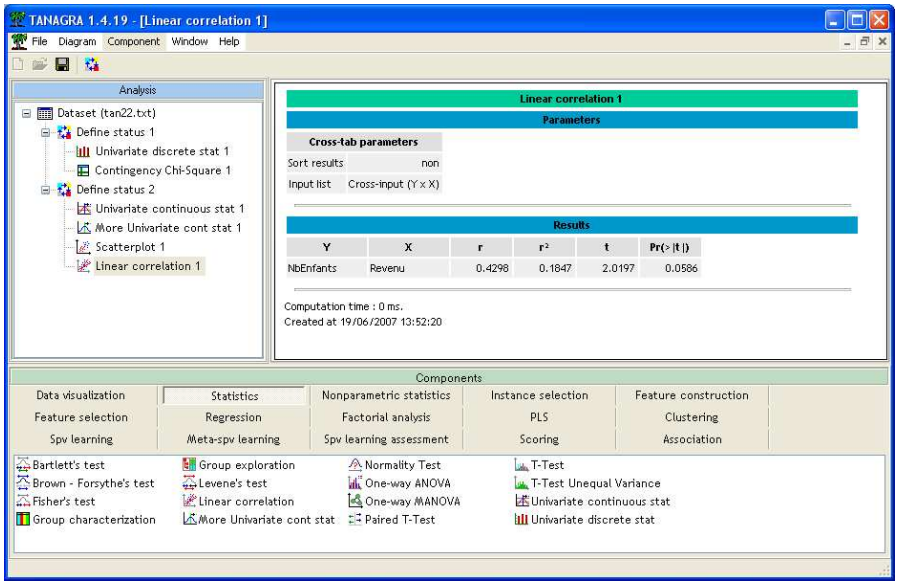
Indicateurs numériques

Pour analyser la relation entre deux variables continues, l'outil graphique est à privilégier car le lien peut receler des situations très complexes. Il existe néanmoins des indicateurs numériques qui peuvent être intéressants lorsqu'un grand nombre de variables est en jeu.

Parmi les outils disponibles figure le coefficient de corrélation (LINEAR CORRELATION, onglet STATISTICS). Nous l'insérons à la suite du DEFINE STATUS 2, nous indiquons dans le paramétrage que les variables sont toutes en INPUT.



Les résultats montrent alors que le lien n'est pas à négliger, le coefficient de corrélation est égal à 0.4298. Plus ce coefficient s'éloigne de 0 en valeur absolue, plus forte est la liaison linéaire. La valeur maximale est 1 (en valeur absolue).

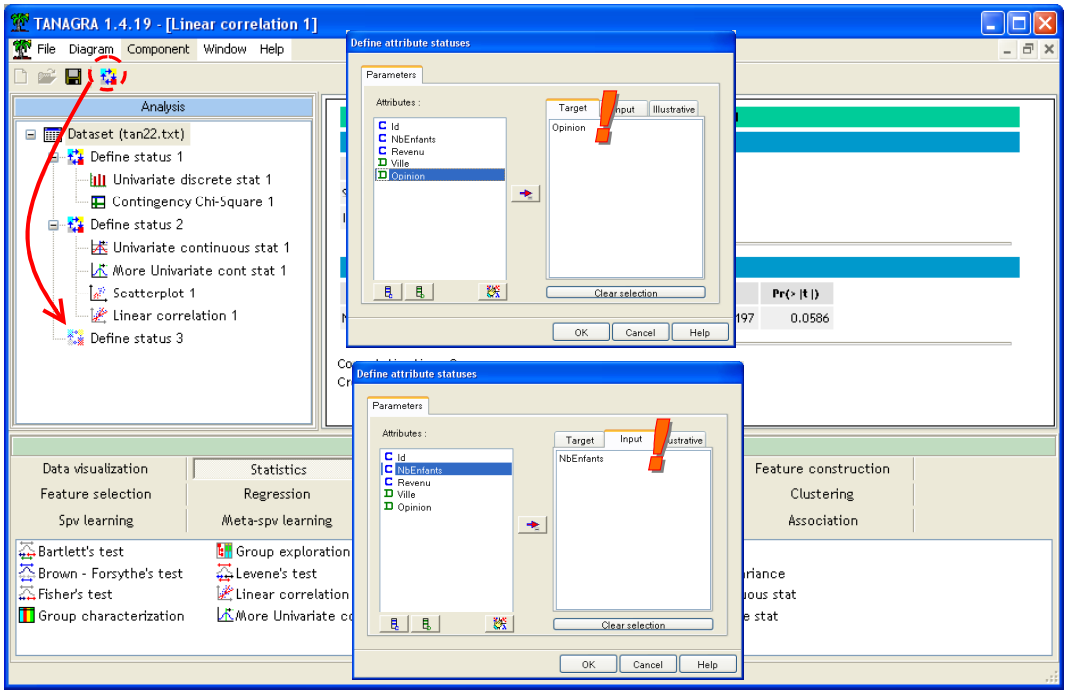


Si la liaison n'est pas linéaire mais reste monotone, nous pouvons utiliser les coefficients de SPEARMAN et KENDALL (onglet NONPARAMETRIC STATISTICS).

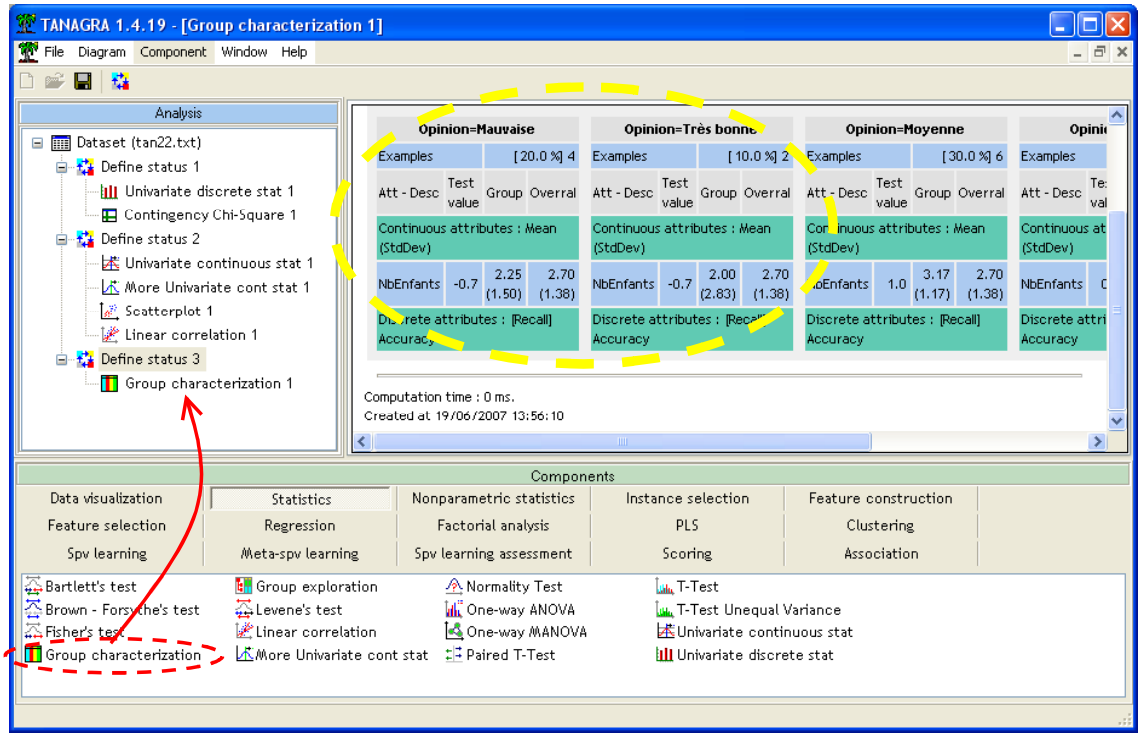
Croisement de variables qualitatives et quantitatives

Croiser des variables quantitatives et des variables qualitatives peut emmener d'autres types d'informations. Dans notre exemple, essayons de comprendre la satisfaction à travers le nombre d'enfant de la personne. Est-ce qu'il y a un lien entre ces deux variables ?

Nous utiliserons l'outil GROUP CHARACTERIZATION pour cela. Insérons tout d'abord un troisième DEFINE STATUS dans le diagramme. Nous plaçons en TARGET la variable OPINION, et en INPUT NBENFANTS.



Nous plaçons alors le composant GROUP CHARACTERIZATION (onglet STATISTICS).



Nous constatons essentiellement que les femmes ayant une opinion très tranchée, bonne ou mauvaise, ont en moyenne moins d'enfants (resp. 2.25 et 2.0) que la population globale (2.7).

Bien entendu, tout ceci n'est qu'un exercice de style, les effectifs sont trop faibles pour conclure à quoique ce soit avec ces données.

Conclusion

Dans la présentation des résultats d'un traitement de données, des indicateurs basiques, des tableaux et des graphiques simples sont au moins aussi souvent pertinents, tout du moins persuasifs, que les méthodes statistiques compliquées, assez obscures pour les non-spécialistes. C'est pour cette raison que les statistiques descriptives tiendront toujours une place prépondérante dans les rapports.