Objectif

Expliciter les formules et les idées véhiculées l'indicateur « valeur test ».

La « valeur test » (VT) est un indicateur qui revient dans plusieurs composants de TANAGRA. Elle sert essentiellement à caractériser un groupe d'individus, défini par une variable caractérisante (ex. le groupe des personnes souffrant d'une maladie), ou issu d'un calcul (ex. classification automatique, un nœud d'un arbre de décision, etc.).

L'objectif est de comprendre ce qui distingue un groupe d'observations à partir d'une série de descripteurs. Pour cela, une procédure simple consiste à comparer les valeurs des paramètres des variables dans la population mère et dans la sous population. Concrètement, lorsque nous manipulons un fichier de données, il s'agit de comparer les valeurs calculées dans l'échantillon initial et dans le sous échantillon correspondant au groupe.

Lorsque la variable est quantitative, nous comparons les moyennes ; lorsqu'elle est qualitative, nous comparons les proportions.

Malgré, ou grâce à sa simplicité, la valeur test est d'un intérêt pratique indéniable. La formulation que nous présentons dans ce didacticiel est tirée de l'ouvrage de Lebart et al. (2000)[1] qui met en avant la VT pour interpréter les groupes issus de la classification automatique, mais aussi pour interpréter les axes factoriels produits par l'analyse des correspondances multiples.

Définitions et formules

Nous disposons d'un échantillon de taille n. Soit un groupe d'observations, définie par la modalité d'une variable quelconque ou produit par une procédure de traitement de donnée, d'effectif n_g. Bien évidemment, $n_g < n$, le sous échantillon est extrait du fichier initial.

Le cas des variables continues

Parmi les variables disponibles, soit **X** une variable quantitative[2]. La moyenne calculée dans l'échantillon global est μ, la variance empirique est égale à σ^2 ; la moyenne calculée dans le sous échantillon est μ_g.

La valeur test (Lebart et al., 2000 ; page 181) est définie de la manière suivante :

[1] L. Lebart, A. Morineau, M. Piron, « Statistique Exploratoire Multidimensionnelle », Dunod, pp.181-184, 2000.

[2] **X** pourrait être également une variable synthétique issue des calculs, c'est le cas des axes factoriels.

$$t_c = \frac{\mu_g - \mu}{\sqrt{\frac{n - n_g}{n-1} \times \frac{\sigma^2}{n_g}}}$$

On distingue au dénominateur l'écart type d'une moyenne dans le cas d'un tirage sans remise de n_g éléments parmi n.

La valeur test t_c peut se lire comme la statistique d'un test de comparaison de moyennes où, sous l'hypothèse nulle de tirage au hasard de n_g parmi n, elle suivrait de manière asymptotique la loi normale centrée réduite. Pour les niveaux de risque usuels (5%), on considérera donc que la différence est significative lorsque la valeur absolue de la VT est supérieure à 2.

Ce n'est pas aussi simple. Dans le cas de la classification, si la variable X a participé à la constitution des groupes, on constate généralement que la valeur test est exagérée. C'était prévisible. Elle a servi à créer des groupes les plus éloignés possibles dans l'espace de représentation, il est normal dans ce cas qu'elle concourt à distinguer les groupes. Dans ce type de processus, on distingue donc les variables actives, qui ont servi au calcul, et les variables illustratives ou supplémentaires, qui sont uniquement mis en avant pour l'interprétation.

Autre écueil important, on constate dans la pratique que la VT prend des valeurs très élevées dès que l'on manipule des bases de taille importante. Laissant à croire que toutes les variables sont significatives dans la distinction des groupes. On comprend mieux ce mécanisme à la lecture de la formule. Si tous les effectifs sont multipliés par un facteur α, la VT est mécaniquement multipliée (approximativement) par un facteur $\sqrt{\alpha}$. Pour un fichier de taille 100 fois plus élevée, avec les mêmes proportions et les mêmes écarts, nous aurions des VT 10 fois plus élevées pour l'ensemble des variables.

Pour ces raisons, il ne faut pas trop s'attacher à comparer la VT avec un hypothétique seuil, très difficile à définir dans la pratique. Il convient avant tout de s'en servir comme un critère permettant de hiérarchiser les variables, afin de distinguer les variables qui jouent un rôle notoire dans l'interprétation des groupes. Il importe surtout de situer les décrochements de valeurs c.-à-d. distinguer les situations où les VT s'écartent fortement des autres.

Le cas des variables catégorielles

Soit Y une variable catégorielle, elle peut prendre, entre autres, la modalité **j**. On note n_j le nombre d'observations portant le caractère j dans l'échantillon global ; n_{jg} dans le sous échantillon, associé au groupe (Y = j), d'effectif n_g.

Si le groupe a été extrait au hasard de l'échantillon initial, l'effectif espéré des individus (Y = j) dans le sous échantillon serait

$$\pi = \frac{n_g \times n_j}{n}$$

La valeur test s' écrit dans ce cas (Lebart et al., 2000 ; page 184) :

$$t_d = \frac{n_{jg} - \frac{n_g \times n_j}{n}}{\sqrt{\frac{n - n_g}{n-1} \times \left(1 - \frac{n_j}{n}\right) \times \frac{n_g \times n_j}{n}}}$$

La distribution asymptotique est gaussienne, on pourrait la comparer avec des seuils prédéfinis. Mais les commentaires émis plus haut – valeurs biaisées pour les variables actives, et surtout sensibilité aux effectifs – font qu' il est encore une fois plus judicieux de repérer les dérochements de valeurs dans les tableaux de résultats.

Dans Tanagra, les résultats (variable = modalité) étant triées selon les VT, on se concentrera avant tout sur le haut et le bas des tableaux.

Un exemple d'application

Pour illustrer notre propos, nous allons effectuer une caractérisation de groupes dans TANAGRA. Concrètement aux didacticiels habituels, nous nous attacherons surtout à décrire attentivement les résultats en mettant en relation les formules ci-dessus et les tableaux de résultats fournis par le logiciel.

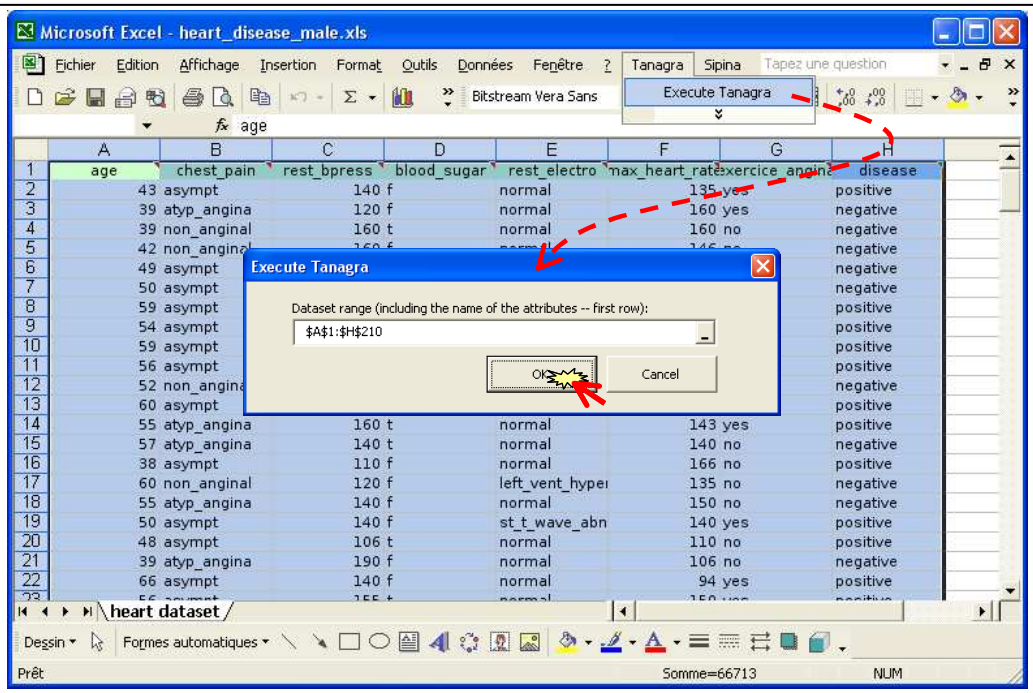
Données

Nous utilisons le fichier HEART_DISEASE_MALE.XLS[3] décrivant des hommes souffrant ou non de maladie cardiovasculaire. L' objectif est de mettre en exergue les caractéristiques des personnes malades.

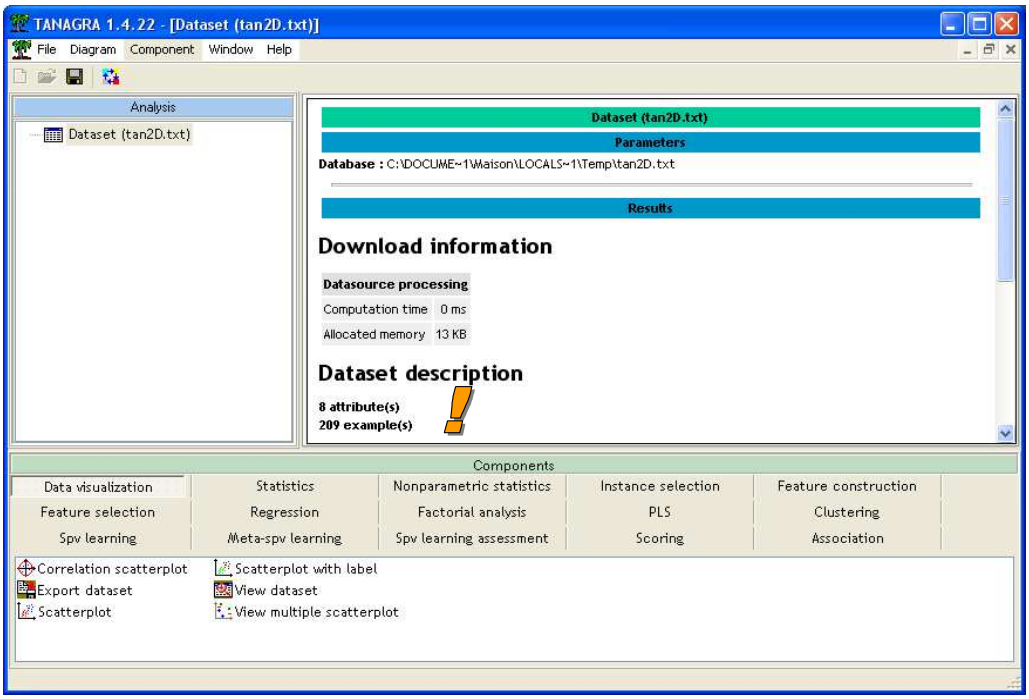
Nous ouvrons le fichier dans le tableur EXCEL. Nous sélectionnons la plage de données, puis nous activons le menu TANAGRA / EXECUTE TANAGRA[4]. Une boîte de dialogue apparaît demandant confirmation de la plage de cellules. Nous validons si la sélection est correcte.

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

[4] Ce nouveau menu dans EXCEL est installé automatiquement à l' aide de la macro complémentaire TANAGRA.XLA, voir http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Excel_AddIn.pdf pour plus de précisions.

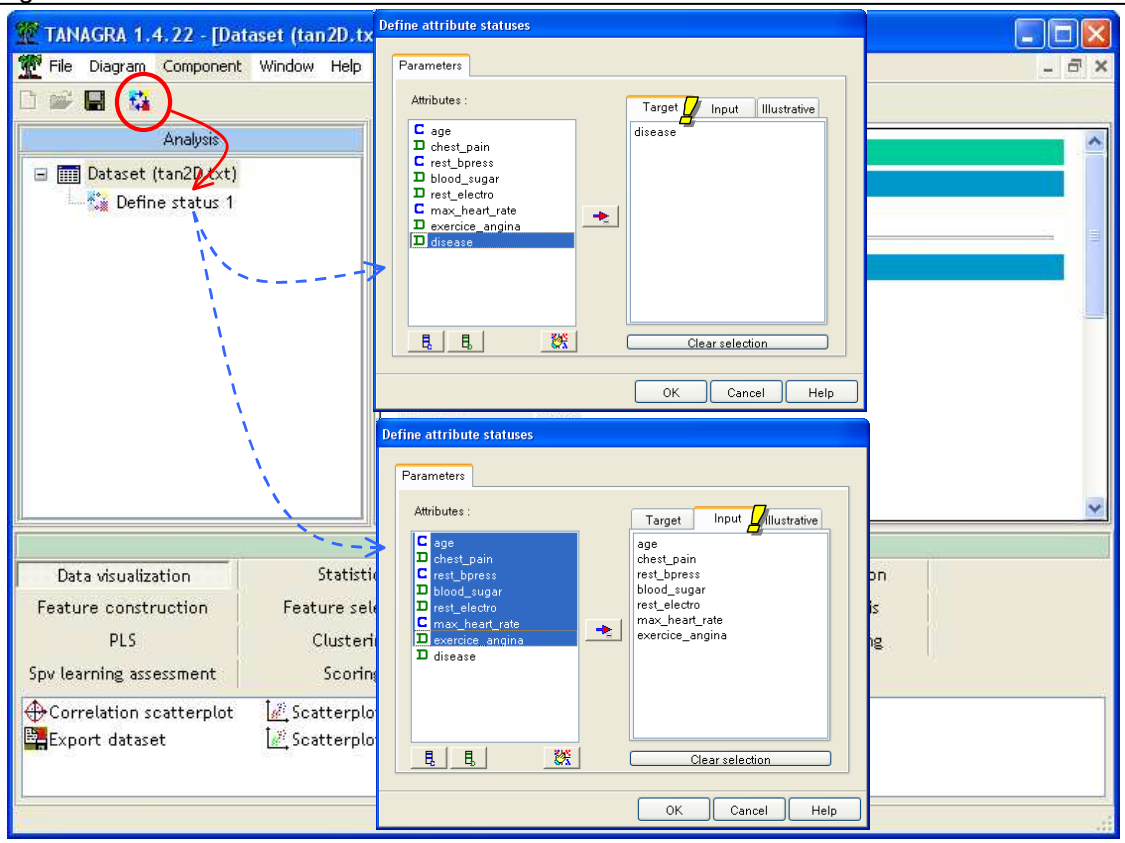


TANAGRA est démarré, les données sont automatiquement chargées. Il y a $n = 209$ observations dans le fichier.



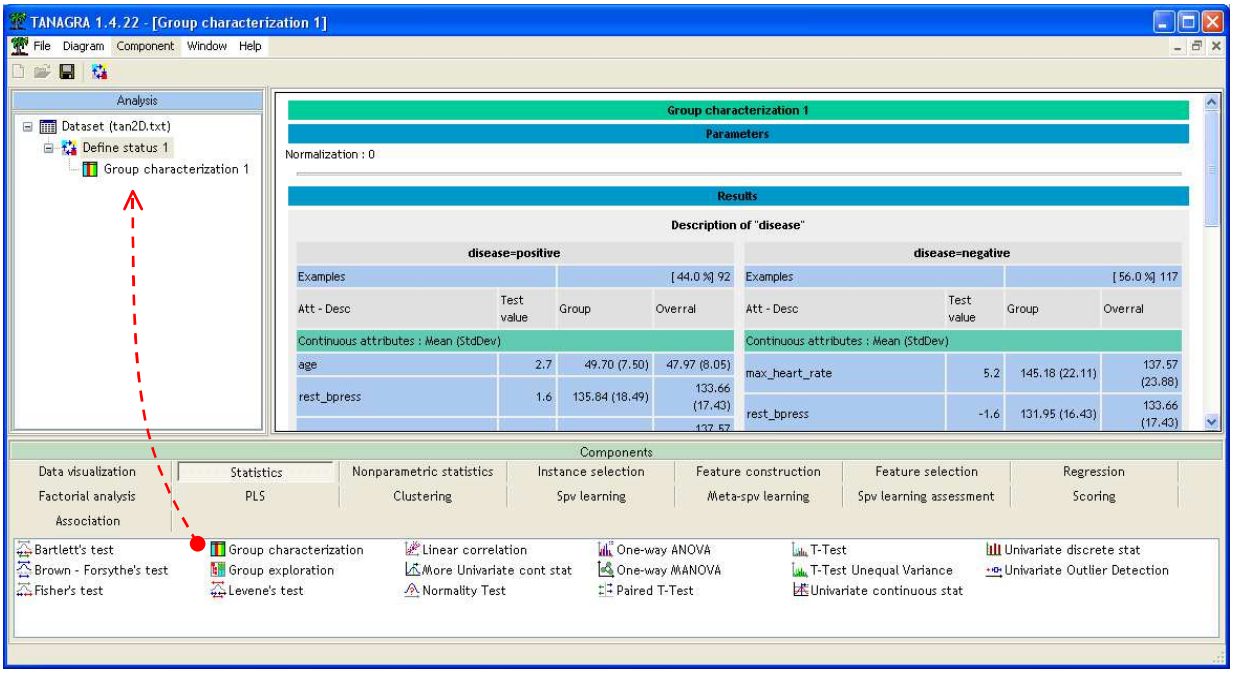
Définir l' analyse

Nous voulons décrire les groupes d' observations correspondant aux malades (DISEASE = POSITIVE) à l' aide des autres variables de la base. Pour ce faire, nous introduisons le composant DEFINE STATUS dans le diagramme, le mieux est de passer par le raccourci dans la barre d' outils. Nous plaçons DISEASE en TARGET, les autres variables en INPUT.



Caractérisation des groupes

Pour caractériser le groupe DISEASE = POSITIVE à l' aide des variables INPUT, nous insérons le composant GROUP CHARACTERIZATION (onglet STATISTICS) dans le diagramme. Nous activons le menu contextuel VIEW, les résultats s' affichent.



L' effectif des personnes malades est $n_g = 92$.

Le tableau de caractérisation est subdivisé en deux parties selon le type des variables. Nous détaillons les résultats pour le groupe DISEASE = POSITIVE.

disease=positive			
Examples			[44.0 %] 92
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
age	2.7	49.70 (7.50)	47.97 (8.05)
rest_bpress	1.6	135.84 (18.49)	133.66 (17.43)
max_heart_rate	-5.2	127.90 (22.61)	137.57 (23.88)
Discrete attributes : [Recall] Accuracy			
chest_pain=asympt	8.4	[73.5 %] 81.5 %	48.80%
exercice_angina=yes	8.3	[83.3 %] 65.2 %	34.40%
blood_sugar=t	2.1	[68.8 %] 12.0 %	7.70%
rest_electro=st_t_wave_abnormality	1.5	[56.7 %] 18.5 %	14.40%
chest_pain=typ_angina	1.1	[66.7 %] 4.3 %	2.90%
rest_electro=?	1.1	[100.0 %] 1.1 %	0.50%
rest_electro=left_vent_hyper	-1.1	[20.0 %] 1.1 %	2.40%
rest_electro=normal	-1.2	[42.2 %] 79.3 %	82.80%
blood_sugar=f	-2.1	[42.0 %] 88.0 %	92.30%
chest_pain=non_anginal	-3.3	[19.4 %] 7.6 %	17.20%
chest_pain=atyp_angina	-6.8	[9.2 %] 6.5 %	31.10%
exercice_angina=no	-8.3	[23.4 %] 34.8 %	65.60%

Variables quantitatives

Attardons nous sur la variable AGE pour le groupe DISEASE = POSITIVE. L'âge moyen est estimé à 47.97 dans la population. Dans le groupe des malades, il serait 49.70. Ces personnes seraient donc plus âgées que le reste de la population. Nous observons entre parenthèses les écarts-types estimés.

Nous pouvons en déduire la valeur test :

$$t_c = \frac{49.70 - 47.97}{\sqrt{\frac{209-92}{209-1} \times \frac{8.05^2}{92}}} \approx 2.74$$

Que l'on retrouve dans le tableau de résultats.

Les personnes sont donc plus âgées. Elles ont un MAX HEART RATE (maximum de pulsations atteint) plus faible (VT = -5.2).

Variables catégorielles

La première modalité qui caractérise les personnes malades est CHEST PAIN = ASYMPT. Nous observons dans le tableau : dans l'échantillon global, 48.8% des personnes présentent CHEST PAIN = ASYMPT ; dans le groupe des malades, la proportion passe à 81.5% ; de plus, 73.5% des personnes CHEST PAIN = ASYMPT se retrouvent dans ce groupe.

Remarque : On observe une double attraction des caractéristiques ici. Ce n' est pas toujours vrai. Prenons le cas de la modalité BLOOD SUGAR = T. La proportion dans la population est 7.7%, elle est de 12% parmi les malades. En revanche, on constate que 68.8% des personnes ayant un fort taux de sucre dans le sang présentent une maladie cardiaque : avoir du sucre dans le sang rend malade, mais être malade ne veut pas dire forcément qu' on a du sucre dans le sang. La distinction est importante.

Revenons à CHEST PAIN = ASYMPT, détaillons les calculs : $n = 209$, $n_g = 92$, $n_j = 0.488 \times 209 = 102$, $n_{jg} = 0.815 \times 92 = 75$,

$$t_c = \frac{75 - \frac{92 \times 102}{209}}{\sqrt{\frac{209-92}{209-1} \times \left(1 - \frac{102}{209}\right) \times \frac{92 \times 102}{209}}} \approx 8.37$$

Pour les variables catégorielles, nous constatons que les caractéristiques CHEST PAIN = ASYMPT, EXERCICE ANGINA = YES, CHEST PAIN = ATYP ANGINA, et EXERCICE ANGINA = NO se démarquent des autres.

Conclusion

La caractérisation des groupes est une tâche clé de l' exploration des données. L' analyse univariée, basée sur des comparaisons de moyennes ou de fréquences, aussi élémentaire soit-elle, donne souvent des indications très précieuses. L' indicateur « valeur test » nous offre la possibilité de hiérarchiser les variables selon leur pertinence.