

Objectif

Mesurer l'importance de la relation entre variables ordinales.

La manipulation d'une variable ordinale n'est pas facile. D'un côté, il s'agit bien d'une variable qualitative, le nombre de valeurs qu'elle peut prendre est réduit. Mais à la différence des variables nominales, les modalités sont ordonnées. De l'autre, nous ne pouvons pas l'assimiler à une variable quantitative, l'amplitude des écarts n'est pas quantifiable. Il faudra tenir compte de ces contradictions lors du choix des outils destinés à extraire de l'information à partir de ces données.

Dans ce didacticiel, nous étudions la mise en œuvre de quelques mesures de dépendances entre 2 variables ordinales (Y et X) dans TANAGRA (Version 1.4.19). Ce document constitue le contrepoint d'un précédent didacticiel traitant des variables nominales¹. Le point de départ est toujours le tableau de contingence croisant les deux variables à analyser. Nous en extrayons les informations nécessaires à la construction des indicateurs. L'essence des mesures est en revanche complètement différente. Elles s'appuient sur la notion de comparaison par paires. On dit qu'une paire d'observations i et j est

- Concordante si $x_i > x_j$ alors $y_i > y_j$;
- Discordante si $x_i > x_j$ alors $y_i < y_j$;
- Dans les autres cas, il y a un ex-aequo sur X ($x_i = x_j$) et/ou sur Y ($y_i = y_j$).

Nous définirons la variable Y comme la variable dépendante (expliquée) et X la variable indépendante (explicative). C'est important pour la lecture et l'interprétation des résultats, mais sans aucune incidence sur les calculs lorsque les mesures sont symétriques c.-à-d. lorsqu'elles fournissent la même valeur quand bien même les variables seraient interverties dans le tableau de contingence. Seul le d de Sommers, qui est une mesure asymétrique, tiendra compte explicitement du rôle des variables.

Bien sûr, de par leur nature qualitative, nous pouvons toujours utiliser les indicateurs dédiés aux variables nominales telles que le KHI-2 ou le U de Theil. Mais dans ce cas, nous mettons de côté le caractère ordinal des modalités. Nous pouvons seulement prédire, s'agissant des mesures asymétriques, si X prend telle valeur, alors Y a de bonnes probabilités de prendre telle valeur. Nous ne répondons pas à la question, lorsque X a tendance à prendre des valeurs élevées, Y aura alors tendance à prendre également des valeurs élevées (liaison positive) ou faibles (liaison négatives). Cette information, qui peut être importante pour l'étude menée, est totalement ignorée.

Bien sûr également, de par leur nature ordonnée, nous pourrions utiliser les techniques largement répandues dédiées aux variables quantitatives telles que l'analyse de corrélation. Mais dans ce cas, les résultats deviennent dépendants du codage adopté. Si nous choisissons de coder la première modalité 0, la seconde 1, la troisième 2, nous obtiendrons un coefficient totalement différent de ce que l'on aurait obtenu avec le triplet (0, 28, 30). La raison est que le coefficient de corrélation tient compte de l'échelle des valeurs, échelle qui est difficile à définir lorsqu'il s'agit de manipuler une variable ordinale. Par exemple, l'écart entre un client mécontent et un client satisfait est-il le même qu'entre un client satisfait et un client très satisfait ? Plutôt que de se perdre dans la recherche d'un codage optimal des modalités. Nous pourrions nous tourner vers les techniques fondées sur les rangs telles que le coefficient de Spearman, qui n'est, ni plus ni moins, que le coefficient de corrélation calculé sur les rangs des observations sur chaque variable. Déjà nous nous

¹ http://eric.univ-yon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Measures_of_Association_Nominal_Variables.pdf

affranchissons de l'échelle des valeurs dans ce cas, c'est plutôt positif. Malheureusement apparaît alors un inconvénient qui peut gêner considérablement les calculs, le nombre de modalités étant relativement faible, il y a de très nombreux ex-aequo. Le coefficient de Spearman, il en est de même pour les autres techniques fondées sur les rangs, les appréhendent mal.

Ceci est la théorie. Dans la pratique, j'ai souvent constaté que le coefficient de corrélation sur des données codées de manière très fruste (codage 1, 2, 3, etc.) donne quand même des indications assez bonnes sur la réalité et la force de la liaison entre deux variables ordinales. On peut le calculer dans un premier temps pour défricher les données. Il nous appartient par la suite d'utiliser les techniques adéquates pour confirmer ces premières impressions, et surtout produire des résultats statistiquement valables et interprétables.

Pour ceux qui veulent approfondir l'étude des mesures présentées dans ce didacticiel, nous citerons 3 références qui réunissent, en grande partie, les indications nécessaires à leur compréhension :

- http://eric.univ-yon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Measures_of_Association_Nominal_Variables.pdf
- <http://www2.chass.ncsu.edu/garson/PA765/assocordinal.htm>
- <http://v8doc.sas.com/sashtml/stat/chap28/sect20.htm>

Données

Les données ([blood_pressure_ordinal_association.xls](#)) utilisées dans ce didacticiel proviennent d'une étude de cas publiée sur le web². L'objectif est d'expliquer l'hypertension artérielle des patients à partir de leurs caractéristiques physiologiques, cliniques et comportementales : le sexe, fumer ou pas, effectuer régulièrement des exercices physiques, les marqueurs génétiques, etc.

Variables

La variable dépendante est au départ la pression systolique (SYSTOLIC). Nous avons décidé de la découper en 3 intervalles (BP3Levels) en nous appuyant sur les valeurs limites couramment utilisées dans le domaine³ :

- tension normale si PA systolique ≤ 140 mm hg
- hypertension élevée si PA systolique > 140 mm g et ≤ 180 mm hg
- hypertension sévère si PA systolique > 180 mm hg

Remarque : Découpage des variables continues. Un autre découpage en 2 modalités a été évalué (BP2Levels). Nous y reviendrons à la fin de ce document. Nous constaterons notamment que lorsque nous découpons une variable quantitative, le choix du nombre et des bornes des intervalles n'est pas innocent. Il peut influencer les résultats au point de conduire à des conclusions contradictoires.

² <http://www.math.yorku.ca/Who/Faculty/Ng/ssc2003/BPMain.htm>

³ <http://www.infirmiers.com/etud/cours/cardio/Htacardio.php>

Pour expliquer l'hypertension, nous avons retenu 9 variables indépendantes :

Variables indépendantes	Description
Gender_M	Sexe (1 : masculin ; 0 : féminin)
Smoke_Y	Fumeur (1 : oui ; 0 : non)
Exercise	Niveau d'activité physique (1 : faible ; 2 : moyen ; 3 : élevé)
Overweight	Corpulence (1 : normal ; 2 : surcharge pondérale ; 3 : obèse)
Alcohol	Consommation d'alcool (1 : faible ; 2 moyenne ; 3 : élevée)
Stress	Niveau de stress (1 : faible ; 2 : moyen ; 3 : élevé)
Salt	Consommation de sel (1 : faible ; 2 : moyen ; 3 : élevée)
Income	Niveau de revenu (1 : faible ; 2 : moyen ; 3 : élevé)
Education	Niveau d'éducation (1 : faible ; 2 : moyen ; 3 : élevé)

Observations

Le fichier original comprenait des individus ayant été traité contre l'hypertension, une variable indicatrice permettait de les distinguer. De fait, cette dernière masquait toutes les informations intéressantes. Les individus non malades se différenciaient avant tout parce qu'ils ont été traités. Nous avons donc décidé de ne conserver que les individus qui n'ont pas été soignés. Le fichier ainsi formé comporte **399 observations**.

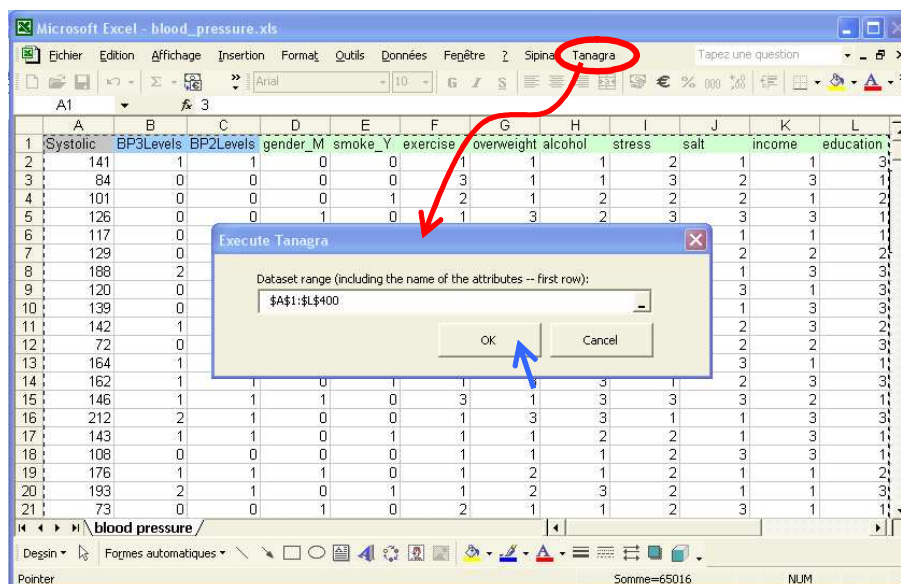
Dans la copie d'écran suivante, nous distinguons les 20 premières observations de notre fichier.

	A	B	C	D	E	F	G	H	I	J	K	L
	Systolic	BP3Levels	BP2Levels	gender_M	smoke_Y	exercise	overweight	alcohol	stress	salt	income	education
2	141	1	1	0	0	1	1	1	2	1	1	3
3	84	0	0	0	0	3	1	1	3	2	3	1
4	101	0	0	0	1	2	1	2	2	2	1	2
5	126	0	0	1	0	1	3	2	3	3	3	1
6	117	0	0	0	1	3	1	2	2	1	1	1
7	129	0	0	0	1	2	1	3	1	2	2	2
8	188	2	1	0	0	2	3	3	1	1	3	3
9	120	0	0	0	0	3	1	2	1	3	1	3
10	139	0	0	0	0	2	1	1	3	1	3	3
11	142	1	1	0	0	1	1	3	3	2	3	2
12	72	0	0	1	0	3	1	1	3	2	2	3
13	164	1	1	1	1	1	1	3	2	3	1	1
14	162	1	1	0	1	1	3	3	1	2	3	3
15	146	1	1	1	0	3	1	3	3	3	2	1
16	212	2	1	0	0	1	3	3	1	1	3	3
17	143	1	1	0	1	1	1	2	2	1	3	1
18	108	0	0	0	0	1	1	1	2	3	3	1
19	176	1	1	1	0	1	2	1	2	1	1	2
20	193	2	1	0	1	1	2	3	2	1	1	3
21	73	0	0	1	0	2	1	1	2	3	1	1

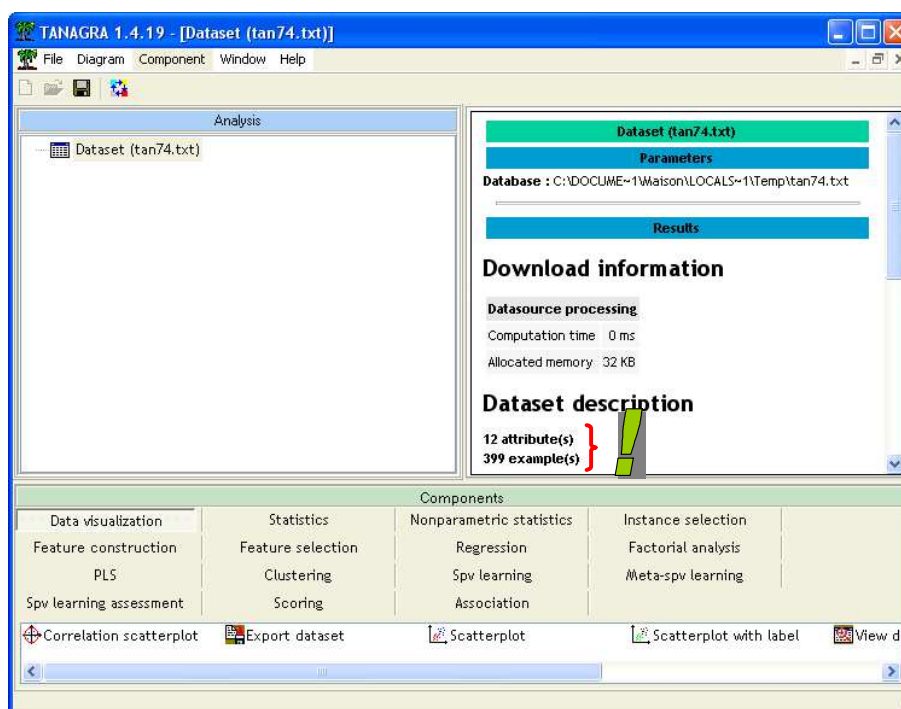
Mesures d'association pour variables ordinales

Importer les données et créer un diagramme

Le plus simple est de charger les données dans le tableur EXCEL. Nous sélectionnons les données, puis, nous lançons TANAGRA en activant le menu TANAGRA/EXECUTE TANAGRA installé par la macro complémentaire TANAGRA.XLA⁴.



TANAGRA est automatiquement démarré, un nouveau diagramme est créé et les données chargées. Nous vérifions que le fichier comporte bien 399 observations et 12 variables.

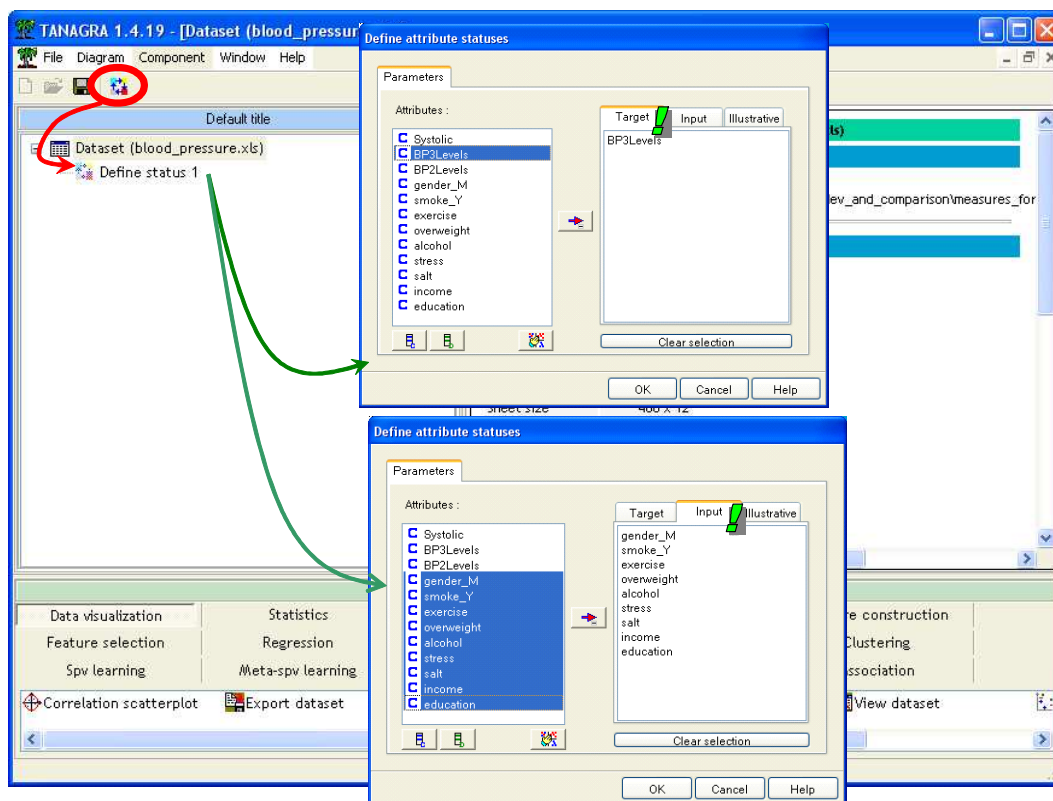


⁴ Cette macro complémentaire est disponible depuis la version 1.4.11 de TANAGRA. Un didacticiel disponible sur le site web indique comment l'activer dans votre tableur EXCEL.

Expliquer les 3 niveaux d'hypertensions

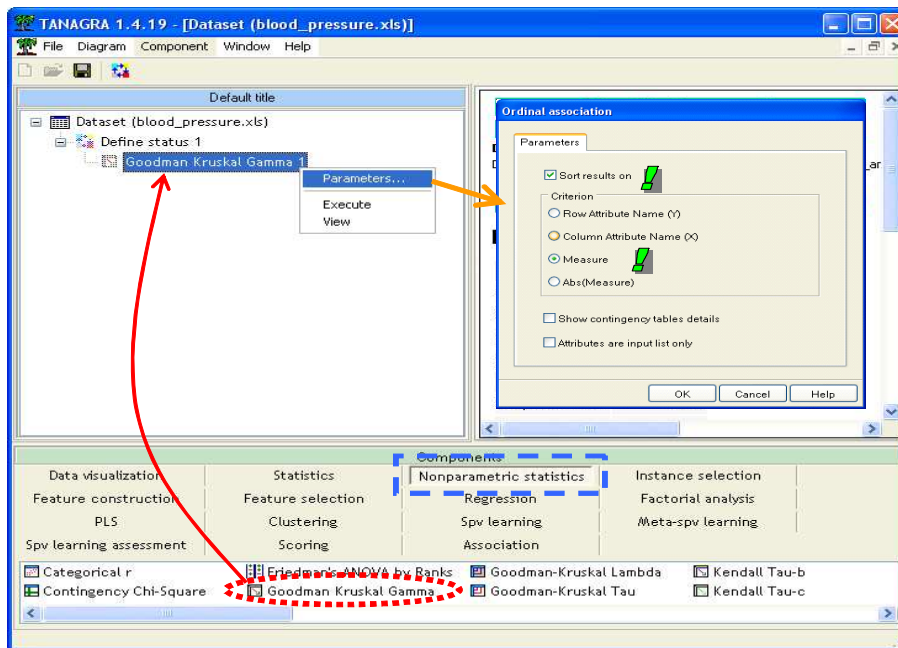
Dans un premier temps, nous étudions le découpage en 3 niveaux de la variable systolique qui nous emmène à distinguer la situation normale et deux niveaux d'hypertension (modérée et sévère). Nous voulons détecter les variables indépendantes qui lui sont liées. Nous utiliserons l'indicateur **Gamma de Goodman et Kruskal**.

Pour ce faire, nous devons tout d'abord définir le rôle des variables à l'aide du composant DEFINE STATUS. Nous l'insérons dans le diagramme en utilisant le raccourci dans la barre d'outils. Nous plaçons en TARGET la variable BP3Levels et en INPUT les variables GENDER_M à EDUCATION.

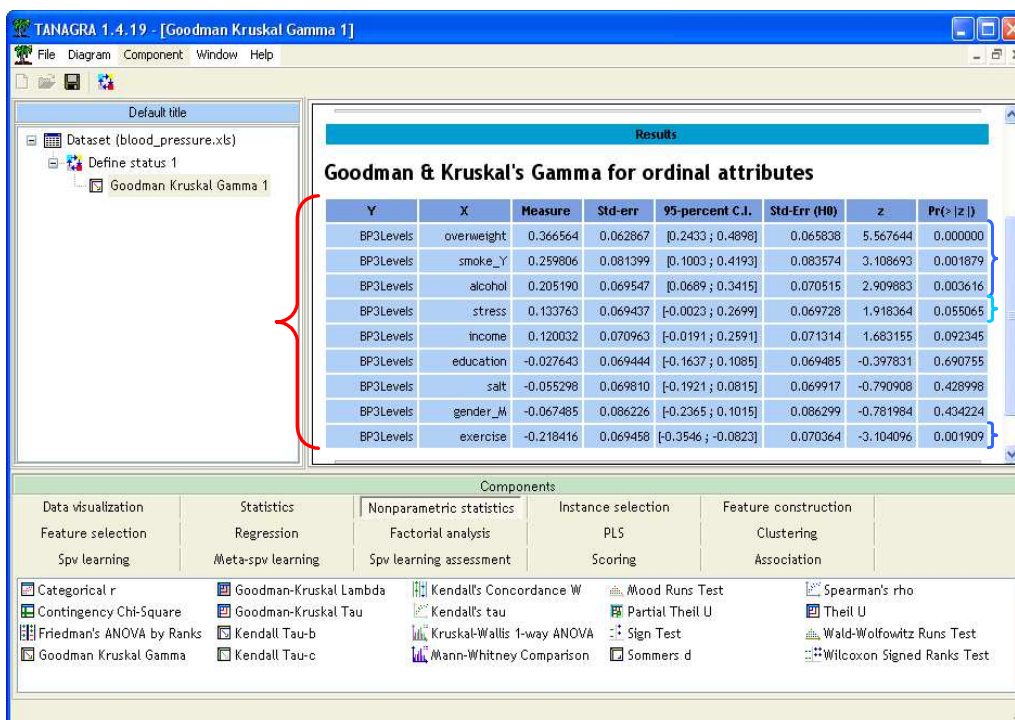


Remarque : codage des variables. Attention ! Les valeurs de la variable BP3Levels correspondent à des codes, comme elles sont ordonnées et surtout puisqu'il s'agit de valeurs numériques, TANAGRA stocke la variable comme une variable continue (quantitative). Lors de la mise en œuvre des calculs, TANAGRA détecte les valeurs distinctes, les trie et les utilise comme des codes dans l'élaboration des tableaux de contingence servant au calcul des indicateurs. Qu'importe que les valeurs soient des valeurs entières ou réelles, il faut simplement qu'elles soient discernables et que l'ordre ait un sens. Si nous voulons utiliser une variable avec des valeurs alphanumériques, par exemple la « qualité » codée « médiocre », « correcte », « très bonne », il nous appartient de transformer manuellement les valeurs en attribuant les codes *avant l'importation* dans TANAGRA (ex. médiocre = 1, correcte = 2, très bonne = 3).

Dans notre diagramme, après le composant DEFINE STATUS, nous insérons le composant Goodman Kruskal Gamma (onglet NONPARAMETRIC STATISTICS). Nous activons le menu PARAMETERS de manière à ce que les résultats soient triés selon l'importance du lien (en premier lieu les variables fortement liées positivement, en dernier les variables liées négativement).



En activant le menu VIEW, nous accédons aux résultats.



Chaque ligne correspond au croisement de 2 variables, nous lisons tour à tour : le nom des variables Y et X, l'indicateur, son écart-type, l'intervalle de confiance à 95%, l'écart type sous hypothèse d'indépendance, la statistique réduite pour le test d'indépendance, et la probabilité critique associée.

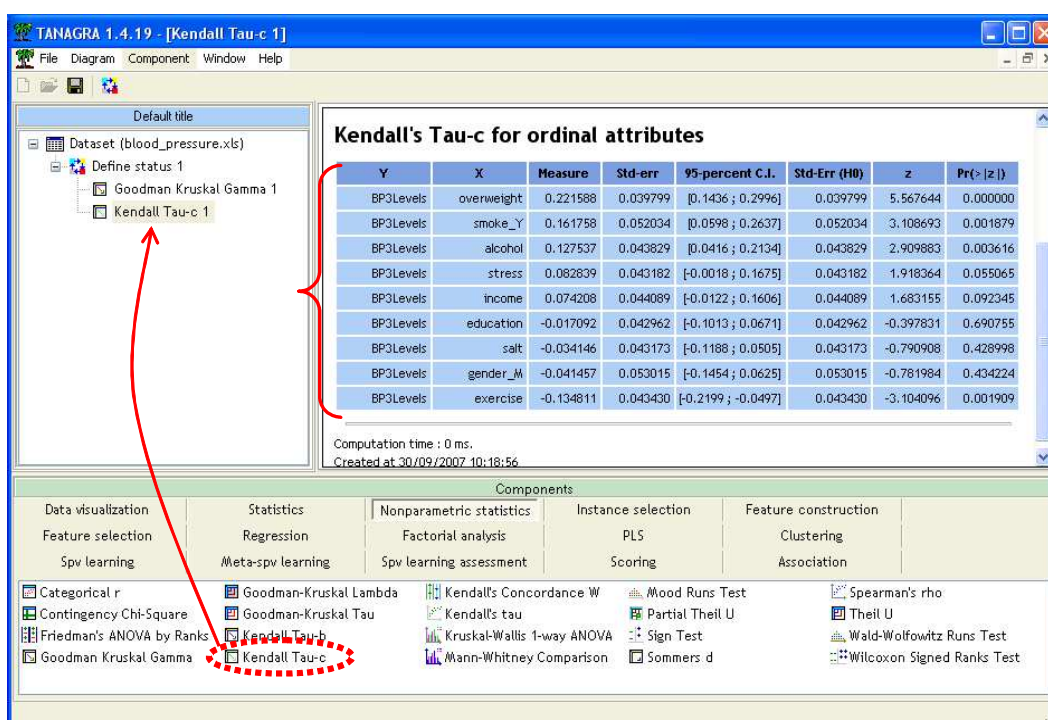
Dans notre exemple, les variables OVERWEIGHT, SMOKE_Y et ALCOHOL sont très significativement (positivement) liées avec l'hypertension. Pour la variable STRESS, la situation est plus contrastée, elle est significative à 10% mais pas à 5%. Pourtant, on sait par ailleurs que le stress n'aide pas à une circulation sanguine apaisée, ce résultat peut poser question, nous y reviendrons à la fin de ce didacticiel. A l'opposé, la variable EXERCISE est négativement liée avec l'hypertension. Comme

quoi se bouger les miches le dimanche matin au parc ne peut que faire du bien. Les autres variables ne semblent pas en relation avec l’hypertension.

Comment interpréter la valeur du coefficient Gamma⁵ ? Voyons le cas de la variable OVERWEIGHT, avec Gamma = 0.367. Nous lirons : pour deux individus pris au hasard, non ex-aequo ni sur X ni sur Y, l’écart entre la probabilité d’être concordant et la probabilité d’être discordant est positif, il est estimé à 0.367. En d’autres termes, lorsque la corpulence augmente, il est probable qu’il en soit de même pour l’hypertension.

Les autres indicateurs symétriques

Nous avons utilisé le Gamma de Goodman et Kruskal pour illustrer l’utilisation des mesures ordinales dans TANAGRA. Nous aurions tout aussi bien pu mettre en œuvre d’autres mesures, avec des conclusions identiques : TAU-B de KENDALL, TAU-C de KENDALL. Lorsque les tableaux sont de taille quelconque, carrée ou rectangulaire, on préférera le TAU-C. Dans notre exemple, le TAU-C fournit le tableau de résultats suivant, toujours en triant les résultats.



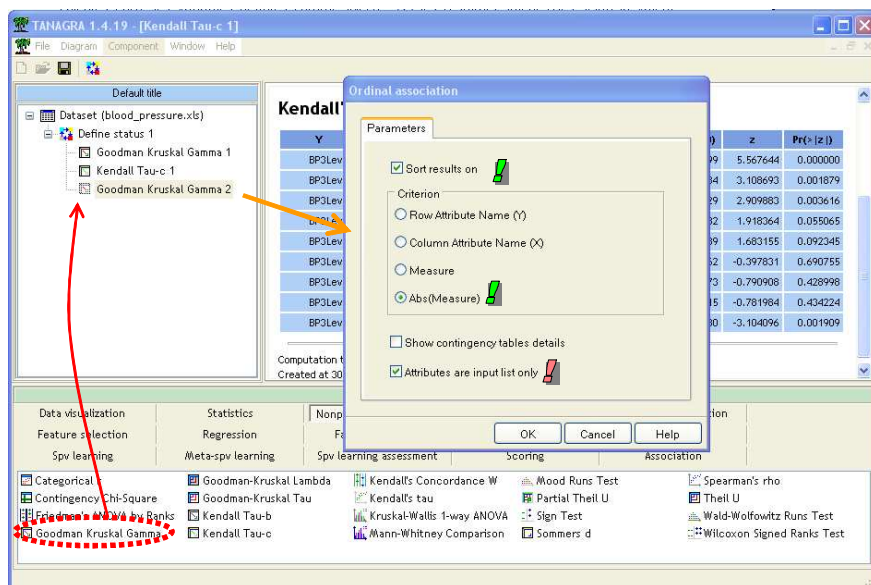
Si les valeurs de l’indicateur sont différentes, les conclusions, et notamment les probabilités critiques des tests de significativité, sont les mêmes.

Associations croisées entre les variables INPUT

Nous avons pu détecter plusieurs variables liées avec l’hypertension. Mais il se peut que ces variables soient par ailleurs redondantes, emmenant le même type d’information sur la variable dépendante. Il se peut également, et c’est plus ennuyeux, que nous ayons des facteurs confondants (Z) qui perturbent la liaison entre deux variables X et Y. Ils peuvent masquer la liaison comme ils peuvent mettre en avant une liaison qui est en réalité fallacieuse. Une bonne manière d’identifier ces situations est de calculer les associations croisées entre les variables INPUT. Nous insérons de nouveau le composant Gamma de Goodman et Kruskal dans le diagramme, nous

⁵ Voir <http://www2.chass.ncsu.edu/garson/PA765/assocordinal.htm>

modifions le paramétrage de manière à ce que : (1) les associations soient calculés entre les variables définies comme INPUT ; (2) les résultats soient triés selon la valeur absolue de la liaison, le nombre de croisements pouvant être très important.



Le détail des résultats est dans le tableau suivant :

Goodman & Kruskal's Gamma for ordinal attributes

Y	X	Measure	Std-err	95-percent C.I.	Std-Err (HO)	z	Pr(> z)
smoke_Y	income	-0.224944	0.078479	[-0.3788 ; -0.0711]	0.080014	-2.811288	0.004934
gender_M	education	-0.192711	0.079493	[-0.3485 ; -0.0369]	0.080652	-2.389418	0.016875
gender_M	alcohol	-0.192669	0.079973	[-0.3494 ; -0.0359]	0.081095	-2.375841	0.017509
smoke_Y	overweight	0.163125	0.082	[0.0024 ; 0.3238]	0.08293	1.967025	0.04918
gender_M	smoke_Y	-0.160004	0.098273	[-0.3526 ; 0.0326]	0.099602	-1.606433	0.108179
salt	education	-0.157149	0.066271	[-0.2870 ; -0.0273]	0.066742	-2.354576	0.018544
overweight	alcohol	-0.146367	0.066696	[-0.2771 ; -0.0156]	0.067109	-2.181032	0.029181
smoke_Y	exercise	0.128124	0.080993	[-0.0306 ; 0.2869]	0.081508	1.571919	0.115969
stress	income	0.104521	0.065635	[-0.0241 ; 0.2332]	0.065841	1.587474	0.112405
exercise	salt	0.092326	0.067841	[-0.0406 ; 0.2253]	0.06802	1.357329	0.174677
gender_M	income	0.089839	0.081352	[-0.0696 ; 0.2493]	0.081617	1.100739	0.27101
alcohol	salt	-0.083667	0.067707	[-0.2164 ; 0.0490]	0.067859	-1.23296	0.217591
gender_M	exercise	-0.083594	0.081649	[-0.2436 ; 0.0764]	0.081857	-1.021213	0.307153
exercise	income	0.074749	0.065819	[-0.0543 ; 0.2038]	0.065926	1.133836	0.256863
smoke_Y	stress	0.071663	0.081441	[-0.0880 ; 0.2313]	0.081601	0.878209	0.37983
gender_M	stress	0.060967	0.082059	[-0.0999 ; 0.2218]	0.082165	0.742014	0.458079
alcohol	income	0.058377	0.068778	[-0.0764 ; 0.1932]	0.068841	0.848003	0.396436
smoke_Y	salt	-0.057968	0.081711	[-0.2181 ; 0.1022]	0.081808	-0.708581	0.478584
income	education	-0.054395	0.067248	[-0.1862 ; 0.0774]	0.067309	-0.808139	0.419011
stress	salt	-0.051508	0.067422	[-0.1837 ; 0.0806]	0.067474	-0.763384	0.445235
smoke_Y	alcohol	-0.05094	0.081703	[-0.2111 ; 0.1092]	0.081782	-0.622872	0.533369
smoke_Y	education	-0.044758	0.081714	[-0.2049 ; 0.1154]	0.081776	-0.547326	0.584155
exercise	education	-0.043204	0.067827	[-0.1761 ; 0.0897]	0.067864	-0.636622	0.524371
alcohol	education	-0.030246	0.066611	[-0.1608 ; 0.1003]	0.066627	-0.453965	0.649854
gender_M	salt	-0.02913	0.08211	[-0.1901 ; 0.1318]	0.082142	-0.354629	0.722868
overweight	salt	-0.028534	0.06827	[-0.1623 ; 0.1053]	0.068285	-0.417873	0.67604
overweight	income	-0.027911	0.067649	[-0.1605 ; 0.1047]	0.067666	-0.412481	0.679987
exercise	alcohol	-0.027764	0.067382	[-0.1598 ; 0.1043]	0.067401	-0.411916	0.680401
exercise	overweight	0.01409	0.068219	[-0.1196 ; 0.1478]	0.068222	0.206525	0.836381
overweight	stress	0.013756	0.069394	[-0.1223 ; 0.1498]	0.069398	0.198219	0.842873
gender_M	overweight	0.01308	0.08425	[-0.1520 ; 0.1782]	0.084257	0.155235	0.876636
overweight	education	0.010495	0.068828	[-0.1244 ; 0.1454]	0.068829	0.152474	0.878813
stress	education	0.009795	0.066509	[-0.1206 ; 0.1401]	0.06651	0.147271	0.882918
exercise	stress	0.007457	0.066921	[-0.1237 ; 0.1386]	0.066921	0.111426	0.911279
alcohol	stress	0.00543	0.067773	[-0.1274 ; 0.1383]	0.067773	0.080118	0.936143
salt	income	0.001247	0.066938	[-0.1299 ; 0.1324]	0.066938	0.018635	0.985132

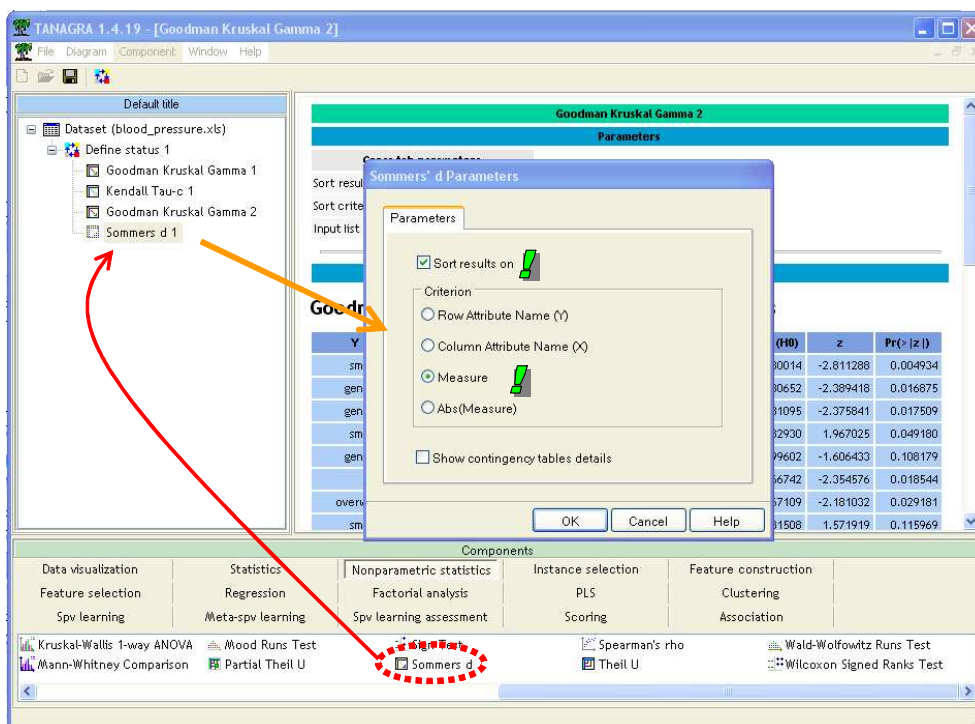
Nous nous en tiendrons aux liaisons significatives à 5%. Le niveau de revenu et le fait de fumer sont négativement liés (-0.224944). Moins on a de l'argent, plus on le fait partir en fumée. D'autres liaisons négatives sont également mis à jour : les hommes sont peu éduqués (à force de regarder le foot je pense, -0.192711), il sont, comme les personnes en surpoids, plus enclins à la tempérance (ah bon ? -0.192669 et -0.146367) ; les personne éduquées ne se gobergent pas de sel (-0.157149). Du côté des associations positives, on constate que le surpoids et la cigarette sont liés positivement.

Plaisanterie mise à part, je pense surtout qu'il faut bien connaître le domaine pour apprécier pleinement ces résultats, voire y détecter les artefacts. Ce n'est pas mon cas.

Association asymétrique

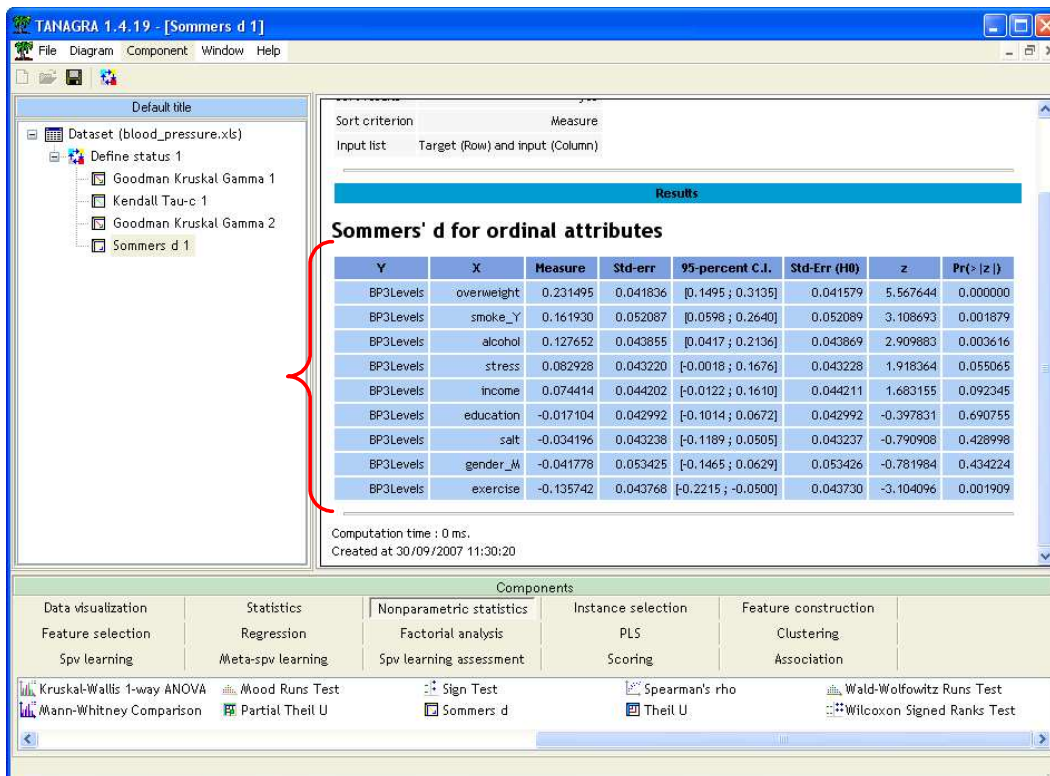
Jusqu'à présent, dans notre esprit, l'idée était d'expliquer l'hypertension à l'aide d'autres variables. Dans les faits, nous évaluons uniquement la liaison sans qu'elle soit directionnelle. Que l'hypertension BP3Levels soit en ligne ou en colonne dans le tableau de contingence ne modifie pas les calculs. Dans cette section, nous introduisons une mesure asymétrique, le d de Sommers. Dans ce cas, on recherche bien une causalité puisque le positionnement de la variable en TARGET ou INPUT pèse sur les résultats.

Nous ajoutons le composant SOMMERS D dans le diagramme. Nous le paramétrons de manière à ce que l'affichage du tableau soit trié selon la valeur de l'indicateur.



Dans notre fichier, le d de Sommers fournit une hiérarchie conforme à ce que proposaient les mesures symétriques. Si nous inversons la sélection TARGET – INPUT, nous n'aurons bien évidemment pas les mêmes valeurs. Certains logiciels proposent également un d de Sommers symétrique.

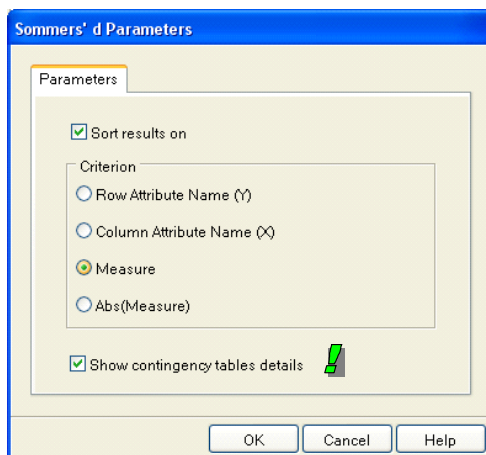
L'interprétation du d de Sommers est un peu différente du Gamma de Goodman et Kruskal. Pour une paire qui n'est pas un ex-aequo sur X (exclusivement), l'écart entre la probabilité d'être concordant et discordant est 0.231495.



Détail des associations – Tableaux de contingence

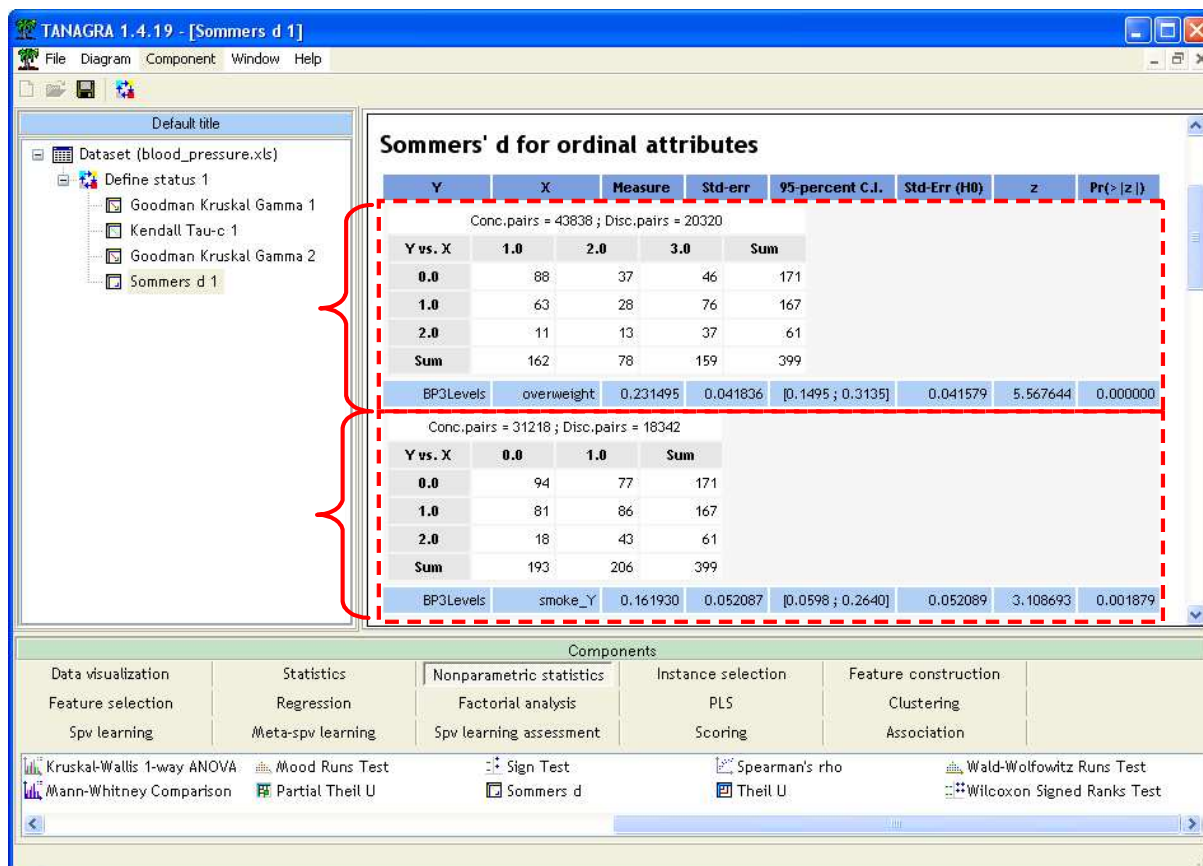
Pour approfondir l'analyse, il peut être intéressant de détailler les tableaux de contingence. L'idée est de caractériser le type de d'association qui peut exister entre les variables⁶ (ex. relation monotone stricte, monotonie prédictive, monotonie faible, etc.).

Pour ce faire, revenons au paramétrage du composant « SOMMERS D 1 » dans notre diagramme (l'option est disponible dans tous les composants mesurant l'association entre variables ordinales). Nous sélectionnons l'option adéquate.



⁶ Voir à ce sujet l'excellent site -- <http://www2.chass.ncsu.edu/garson/PA765/association.htm>

Les indicateurs sont maintenant affichés avec le détail du tableau de contingence croisant les variables. Nous voyons les résultats pour les 2 premières associations (BP3LEVELS vs. OVERWIEGHT et BP3LEVELS vs. SMOKE_Y).



Pour le cas BP3LEVELS vs. OVERWEIGHT, nous lisons : Il y a 43838 paires concordantes, 20320 paires discordantes ; le d de Sommers est égal à 0.231495, il est très fortement significatif.

Si l'on se réfère à la caractérisation de l'association à l'aide du tableau de contingence, nous sommes dans le cadre d'une monotonie assez faible : les maximums des effectifs sont situés sur la première ligne dans la 1^{ère} et 2^{ème} colonne, il est par la suite dans la 2^{ème} ligne pour la 3^{ème} colonne. Par ailleurs, les autres cases sont fortement polluées. Représenter le tableau sous forme de profils colonnes peut être intéressant pour mieux appréhender l'association prédictive.

Influence du codage des variables quantitatives

Parfois, les variables ordinales sont originellement quantitatives, elles sont découpées en intervalles afin de pouvoir les mettre en relation avec d'autres variables, les rendre compatibles avec les mesures utilisées, ou tout simplement parce qu'ainsi nous sommes plus en adéquation avec le contexte de l'étude.

Prenons l'exemple des excès de vitesses et les conséquences qui en découlent. La valeur de la vitesse en elle-même, bien qu'informatrice, est peu intéressante. En revanche, pouvoir la situer par rapport aux seuils institutionnels définissant l'échelle de la répression peut être autrement plus intéressante (20 km/h au dessus de la vitesse autorisée, 40 km/h, etc.).

Dans cette opération dite de discrétisation, le choix du nombre et des bornes de découpage des variables continues ne sont pas sans conséquences sur les résultats. En effet, si elle est réalisée à mauvais escient, il est possible que l'on masque des informations importantes dans les données. A

contrario, si elle est réalisée judicieusement, nous pouvons mieux mettre en lumière les liens qui existent.

Cette question du découpage semble moins se poser lorsque les variables sont nativement ordinales (ex. l'expression de la satisfaction, l'évaluation qualitative d'un produit, etc.). Néanmoins, on pourrait également chercher à définir un regroupement efficace des modalités pour mieux identifier les associations.

Dans notre exemple, nous avons codé la variable dépendante en 3 niveaux en nous inspirant des connaissances du domaine. On peut se demander si un autre codage pourrait emmener d'autres types de résultats. Nous avons ainsi introduit une autre variable (BP2Levels) à deux modalités, elle distingue les hypertendus (> 140 mm hg, codé 1) des autres (codé 0). Nous avons mené une seconde analyse en insérant un nouveau DEFINE STATUS dans le diagramme, nous plaçons en TARGET la nouvelle variable dépendante à 2 modalités, et en INPUT les mêmes variables que précédemment. Nous utilisons le Gamma de Goodman et Kruskal. Les résultats font apparaître un phénomène très intéressant.

The screenshot shows the TANAGRA 1.4.19 interface. The main window displays the results of a Goodman & Kruskal's Gamma analysis for ordinal attributes. The results table is as follows:

Y	X	Measure	Std-err	95-percent C.I.	Std-Err (H0)	z	Pr(> z)
BP2Levels	overweight	0.376732	0.074218	[0.2313 ; 0.5222]	0.079128	4.761050	0.000002
BP2Levels	smoke_Y	0.228012	0.096537	[0.0388 ; 0.4172]	0.099257	2.297190	0.021608
BP2Levels	alcohol	0.206170	0.079353	[0.0506 ; 0.3617]	0.080661	2.556017	0.010588
BP2Levels	stress	0.173544	0.080778	[0.0152 ; 0.3319]	0.081750	2.122848	0.033767
BP2Levels	income	0.135216	0.081628	[-0.0248 ; 0.2952]	0.082140	1.646166	0.099729
BP2Levels	education	0.014437	0.082621	[-0.1475 ; 0.1764]	0.082628	0.174722	0.861298
BP2Levels	salt	-0.034674	0.082599	[-0.1966 ; 0.1272]	0.082627	-0.419644	0.674745
BP2Levels	gender_M	-0.102734	0.100497	[-0.2997 ; 0.0942]	0.101162	-1.015543	0.309847
BP2Levels	exercise	-0.243481	0.078736	[-0.3978 ; -0.0892]	0.080757	-3.014992	0.002570

The components panel at the bottom lists various statistical tests, including Goodman-Kruskal Tau, Kendall Tau-b, Kendall Tau-c, Kendall's Concordance W, Kendall's tau, Kruskal-Wallis 1-way ANOVA, Mann-Whitney Comparison, Mood Runs Test, Partial Theil U, Sign Test, Sommers d, Spearman's rho, Theil U, Wald-Wolfowitz Runs Test, and Wilcoxon Signed Ranks Test.

Nous retrouvons toujours les mêmes variables que pour l'analyse de BP3Levels (OVERWEIGHT, SMOKE_Y, ALCOHOL et EXERCISE). Mais cette fois-ci, la variable explicative STRESS devient significative à 5%.

Cela indiquerait que STRESS permet de discerner l'hypertension chez les patients. En revanche elle est inopérante lorsqu'il s'agit d'expliquer la gradation de l'hypertension. Pour bien apprécier cette nuance, nous reproduisons ici les deux tableaux de contingence.

Conc.pairs = 37260 ; Disc.pairs = 28468				
Y vs. X	1.0	2.0	3.0	Sum
0.0	63	58	50	171
1.0	45	60	62	167
2.0	17	22	22	61
Sum	125	140	134	399

BP3Levels	stress	0.133763	0.069437	[-0.0023 ; 0.2699]	0.069728	1.918364	0.055065
-----------	--------	----------	----------	--------------------	----------	----------	----------

Tab. 1 - Analyse à 3 niveaux, BP3Levels vs. STRESS

Conc.pairs = 30660 ; Disc.pairs = 21592				
Y vs. X	1.0	2.0	3.0	Sum
0.0	63	58	50	171
1.0	62	82	84	228
Sum	125	140	134	399

BP2Levels	stress	0.173544	0.080778	[0.0152 ; 0.3319]	0.081750	2.122848	0.033767
-----------	--------	----------	----------	-------------------	----------	----------	----------

Tab. 2 - Analyse à 2 niveaux, BP2Levels vs. STRESS

En passant à 2 niveaux, le nombre de paires concordantes certes diminue, mais la réduction des paires discordantes est proportionnellement plus élevée. L'indicateur Gamma augmente au point de devenir significatif à 5%.

On pourrait bien sûr développer des stratégies pour mettre en évidence des regroupements « optimaux ». Mais cela dépasse largement le cadre de ce didacticiel. L'idée maîtresse qu'il faut retenir à mon sens est qu'un codage n'est jamais immuable, il nous appartient de les adapter aux objectifs de l'étude. Tout en gardant à l'esprit qu'un résultat, fût-il optimal, n'est rien s'il ne correspond pas aux réalités du domaine ou s'il n'est pas interprétable⁷.

Comparer les résultats avec le coefficient de corrélation

Nous avons affirmé, et nous ne sommes pas les seuls, qu'à cause de la nature ordinale des variables, avec de nombreux ex-aequo, le coefficient de corrélation dévolue à l'étude des dépendances entre variables quantitatives n'est pas adapté. Nous avons voulu vérifier cette assertion en insérant le composant LINEAR CORRELATION (onglet STATISTICS) après DEFINE STATUS 1 dans le diagramme, sur l'analyse avec une variable dépendante à 3 niveaux (BP3Levels). Nous le paramétrons (menu contextuel PARAMETERS) de manière à ce que les croisements soient triés selon les valeurs de « r » décroissants. Nous pouvons ainsi confronter directement les résultats avec ceux du Gamma de Goodman et Kruskal sur les mêmes données.

⁷ « Sans maîtrise, la puissance n'est rien » disait un vieux slogan...

Sort criterion: r statistic
Input list: Target (Y) and input (X)

Y	X	r	r ²	t	Pr(> t)
BP3Levels	overweight	0.2640	0.0697	5.4540	0.0000
BP3Levels	smoke_Y	0.1608	0.0259	3.2457	0.0013
BP3Levels	alcohol	0.1491	0.0222	3.0034	0.0028
BP3Levels	stress	0.0897	0.0080	1.7935	0.0737
BP3Levels	income	0.0842	0.0071	1.6844	0.0929
BP3Levels	education	-0.0287	0.0008	-0.5712	0.5682
BP3Levels	gender_M	-0.0341	0.0012	-0.6808	0.4964
BP3Levels	salt	-0.0444	0.0020	-0.8860	0.3761
BP3Levels	exercise	-0.1518	0.0230	-3.0596	0.0024

Components:

- Data visualization
- Feature selection
- Spv learning
- Statistics
 - Regression
 - Meta-spv learning
 - Linear correlation
 - More Univariate cont stat
 - Normality Test
 - One-way ANOVA
- Nonparametric statistics
 - Factorial analysis
 - Spv learning assessment
- Instance selection
 - PLS
 - Scoring
- Feature construction
 - Clustering
 - Association

Other tests available in the toolbar:

- Bartlett's test
- Brown - Forsythe's test
- Fisher's test
- Group characterization
- Group exploration
- Levene's test
- Linear correlation
- More Univariate cont stat
- Normality Test
- One-way ANOVA
- One-way MANOVA
- Paired T-Test
- T-Test
- T-Test Unequal Variance
- Univariate continuous stat
- Univariate discrete stat
- Univariate Outlier Detection

Sans grande surprise, la hiérarchie des variables est identique, il en est de même pour la significativité du lien. Sans grande surprise car, même s'il est assujéti à une contrainte linéaire, le coefficient de corrélation est faite pour détecter une certaine monotonie dans la relation entre les variables. Sauf codage totalement farfelu des modalités, les coefficients calculés et les tests de significativité associés ne sont pas dépourvus de pertinence. En revanche, l'interprétation du carré du coefficient de corrélation sous forme de variance expliquée est nettement moins appropriée.

Autre élément qui peut expliquer cette convergence, la plupart des variables manipulées dans cette étude sont des variables quantitatives traduites en niveaux c.-à-d. discrétisées. La variable « Corpulence » (OVERWEIGHT) par exemple s'avère être le découpage de l'indice de masse corporelle (BMI - présent dans le fichier initial) par strates; la variable cigarette est vraisemblablement une transcription de la consommation journalière, etc. Le regroupement des valeurs en classes n'aura pas (trop) altéré le lien initial entre les variables. Ainsi, le coefficient de corrélation entre la variable SYSTOLIC et BMI est de 0.3159, à comparer avec 0.2640 lorsque les variables sont découpées en intervalles et codées (1, 2, 3).

Conclusion

Dans ce didacticiel, nous avons montré comment mettre en œuvre les composants mesurant les associations entre variables ordinales. 4 composants ont été introduits dans cette version 1.4.19 de TANAGRA : le Gamma de Goodman et Kruskal, les Tau-b et Tau-c de Kendall, et le d de Sommers.

Pour apprécier au mieux la nature et l'interprétation des résultats, nous conseillons la lecture en parallèle des supports de cours et autres sites web référencés dans l'introduction.