REVUE DE STATISTIQUE APPLIQUÉE

H. MESSATFA

Unification de certains critères d'association par linéarisation et normalisation

Revue de statistique appliquée, tome 36, n° 3 (1988), p. 51-68

<http://www.numdam.org/item?id=RSA_1988__36_3_51_0>

Rev. Statistique Appliquée, 1988, **XXXVI** (3), 51-68

UNIFICATION DE CERTAINS CRITÈRES D'ASSOCIATION PAR LINÉARISATION ET NORMALISATION

H. MESSATFA

Université Paris VI – Centre Scientifique IBM France
36 Avenue Raymond Poincaré 75016 Paris

RÉSUMÉ

Cet article passe en revue quelques mesures d'associations. Nous n'essayons pas ici de faire un panorama complet de la littérature concernant le sujet. Nous proposons une comparaison basée sur les concepts de comparaison par paire et linéarisation. Cette structure clarifie les liens entre les mesures traditionnelles.

Mots clés : *Mesure d'association, Linéarisation, Normalisation et Comparaison par paires.*

ABSTRACT

This paper reviews some measures of association. We will not try to review this literature here. Instead, we develop a much broader class of possible comparison measures based on the concepts of paired comparison and normalization. This structure clarifies links between several traditional indices.

Keywords: *Measure of association, Normalization and Paired comparison.*

1. Introduction

Le propos de cette étude est de montrer comment l'introduction du concept de comparaisons par paires, met en évidence les liens entre plusieurs critères d'associations. Ces critères statistiques calculés à partir de tableaux de contingence, donnent lieu à des formulations mathématiques de complexité quadratique, dans l'approche contingentielle. Nous ne prétendons pas donner, un inventaire global des critères usuels d'association mais nous analyserons ceux qui nous paraissent importants pour notre propos, en particulier nous présentons des critères peu connus mais qui ont une grande utilité comme mesure d'association.

Dans un premier temps, on étudiera les liens qui existent entre :

Le critère E (écart à la moyenne ou écart carré à l'indépendance) et le critère B_k de E.B. Fowlkes et C.L. Mallows.

Le critère I (écart à l'indétermination introduit par F. Marcotorchino) et le critère J de S. Janson et J. Vergelius (cas particulier du critère I).

Le critère B de Belson et le critère S de G. Saporta.

Les critères précédents sont basés sur deux concepts statistiques de comportement des cases d'un tableau de contingence $p \times q$ dont les expressions respectives sont avec des notations évidentes :

l'indépendance statistique :

$$n_{uv} = \frac{n_{u.} n_{.v}}{N} \quad \forall (u, v)$$

l'indétermination statistique :

$$n_{uv} = \frac{n_{u.}}{q} + \frac{n_{.v}}{p} - \frac{N}{pq} \quad \forall (u, v).$$

Un nombre considérable d'indices reposent sur l'indépendance qui est bien connue des statisticiens, en revanche l'indétermination peu utilisée a été introduite dans le critère de la variance résiduelle. Ce dernier concept a été développé par F. Marcotorchino dans [8].

La deuxième partie, basée sur deux autres concepts (la normalisation et la linéarisation) a pour but, de faire une synthèse de plusieurs mesures d'associations.

1.1. Notations

Pour les indices présentés dans ce texte, on part de la configuration suivante : C représente une variable à p modalités (partition à p classes d'un ensemble à N éléments) et Y une variable à q modalités (partition à q classes de ce même ensemble); permettant de fournir le tableau de contingence suivant :

n_{uv} = effectif de la case (u, v)

$n_{u.}$ = nombre d'objets ayant la modalité u de C

$n_{.v}$ = nombre d'objets ayant la modalité v de Y

$$n_{u.} = \sum_{v} n_{uv}$$

$$n_{.v} = \sum_{u} n_{uv}$$

$$N = \sum_{u,v} n_{uv}.$$

1.2. Tableau de comparaisons par paires associé à une partition

Les partitions C et Y données peuvent être représentées chacune par un tableau de comparaisons par paires. Ce tableau sera carré de taille $N \times N$ à valeurs dans $(0, 1)$ et le terme général c_{ij} sera défini par :

$$\begin{cases} c_{ij} = 1 & \text{si } i \text{ et } j \text{ sont dans la même classe de } C \\ c_{ij} = 0 & \text{sinon.} \end{cases}$$

On définit de façon analogue le tableau des y_{ij} associé à Y.

Pour les importantes propriétés du tableau de comparaisons par paires C, voir F. Marcotorchino [10].

1.3. Linéarisation

Des formules liant les éléments du tableau de contingence croissant C et Y, et les éléments des tableaux de comparaisons par paires C et Y associés sont données entre autre par :

$$\sum_{u,v} n_{uv}^2 = \sum_{ij} c_{ij} y_{ij} \tag{1}$$

$$\sum_{u} n_{u.}^2 = \sum_{ij} c_{ij} \tag{2}$$

$$\sum_{v} n_{.v}^2 = \sum_{ij} y_{ij} \tag{3}$$

$$\sum_{u,v} n_{uv} n_{u.} n_{.v} = \left(\sum_{ij} (c_{i.} + c_{.j}) y_{ij}\right) \Big/ 2. \tag{4}.$$

Pour plus de détails sur la démonstration des formules voir [8].

On introduira également les notations suivantes :

$$c_{ij} = \sum_{ij} c_{ij} \; ; \quad c_{i.} = \sum_{j} c_{ij} \; ; \quad c_{.j} = \sum_{i} c_{ij} \; ; \quad \bar{c}_{ij} = 1 - c_{ij} \; ; \quad \bar{y}_{ij} = 1 - y_{ij} \, .$$

Les relations liant les formules contingentielles et les formules en comparaisons par paires nous ont aidé à comprendre les critères étudiés et nous avons pu constater que la plupart des indices d'association que nous présenterons ne sont que les variantes d'une covariance non centrée : $A_1 = \sum_{ij} c_{ij} y_{ij}$ où A_1 est le critère des accords positifs, dans la notation de F. Marcotorchino [8].

Normalisation

Pour les indices d'association non nuls en cas d'indépendance statistique, on va proposer une modification permettant cette annulation. On notera $\hat{\Lambda}$ la valeur de l'indice Λ pour $n_{uv} = \dfrac{n_{u.} n_{.v}}{N}$; alors le nouvel indice Λ_N, défini par $\Lambda_N = \dfrac{\Lambda - \hat{\Lambda}}{\Lambda_{\max} - \hat{\Lambda}}$ s'annule en cas d'indépendance,

$$\begin{cases} \Lambda_N = 0 & \text{si } \Lambda = \hat{\Lambda} \\ \Lambda_N = 1 & \text{si } \Lambda = \Lambda_{\max} \end{cases}$$

Λ_N est une normalisation de Λ. On montrera par la suite que par ce procédé de normalisation de quelques indices d'associations, non nuls en cas d'indépendance

statistique, on retrouve différentes normalisations du critère des accords positifs $A_1 = \sum_{u,v} n_{uv}^2 = \sum_{ij} c_{ij} y_{ij}$ et que la plupart des indices d'associations ne sont qu'une transformation géométrique de ce dernier; c'est-à-dire que ces critères s'écrivent sous la forme $\Omega = \frac{1}{Q} \sum_{ij} (c_{ij} - a)(y_{ij} - b)$ où a, b et Q sont des réels et où Q est une norme choisie pour que Ω appartienne à $[-1, 1]$ ou encore $\Omega = \sum_{ij} (S_{ij} - a)(y_{ij} - b)$ où S_{ij} est une transformation sur c_{ij}.

2. Le critère à l'écart à la moyenne

2.1. Présentation

Le critère que l'on étudiera ici prendra tout son sens dans sa formulation en notations par paires; c'est d'ailleurs à cette formulation qu'il doit son nom d'écart à la moyenne. Il a été étudié, dans sa formulation en paires par S. Chah [2] qui l'interprète comme un critère pondéré d'adéquation d'une partition à une préordonnance (dérivé du critère de W.F. DE LA Vega). Dans sa formulation contingentielle F. Marcotorchino [8] l'interprète comme étant l'écart carré à l'indépendance. Parallèlement ce critère à été étudié par A. Agresti et L. Morey [1] comme un ajustement du critère de Rand (Condorcet); car ce dernier ne s'annule pas en cas d'indépendance statistique. Dans cette étude on donnera d'autres interprétations de cet indice.

2.2. Définitions et propriétés

Dans sa formulation contingentielle le critère s'écrit :

$$E(C,Y) = \sum_{u,v} n_{uv}^2 - \frac{\sum_u n_{u.}^2 \sum_v n_{.v}^2}{N^2}.$$

- $E(C,Y)$ s'annule en cas d'indépendance statistique $n_{uv} = \frac{n_{u.} n_{.v}}{N}$ (mais il s'annule également dans d'autres configurations des n_{uv}).
- $E(C,Y)$ est maximum en cas d'association complète : $p = q, n_{u,v} = n_{u.} = n_{.v}$ si $u = v, n_{u,v} = 0$ si $u \neq v$.
- la valeur maximale de l'indice est alors donnée par l'une ou l'autre des 2 quantités suivantes qui sont égales :

$$E_1 = \sum_u n_{u.}^2 \left(1 - \frac{\sum_u n_{u.}^2}{N^2}\right) \qquad E_2 = \sum_v n_{.v}^2 \left(1 - \frac{\sum_v n_{.v}^2}{N^2}\right).$$

Si on n'est pas dans le cas de l'association complète $E_1 \neq E_2$.

Les formules (1), (2), (3) nous permettent d'écrire :

$$E(C,Y) = \sum_{ij} c_{ij} y_{ij} - \frac{c_{..} y_{..}}{N^2} = \sum_{ij} \left(c_{ij} - \frac{c_{..}}{N^2}\right)\left(y_{ij} - \frac{y_{..}}{N^2}\right) = \sum_{ij} \left(c_{ij} - \frac{c_{..}}{N^2}\right) y_{ij}.$$

Remarque 1

E s'interprète donc comme une covariance (au facteur $\frac{1}{n^2}$ près) entre les vecteurs de composantes C_{ij} et y_{ij}.

Remarque 2

$E(C,Y)$ peut s'écrire sous la forme $E(C,Y) = \sum_{ij}\left(c_{ij} - \frac{c_{..}}{N^2}\right)\left(y_{ij} - a\right)$ quelque soit le réel a indépendant de i et j; car $\sum_{ij}\left(c_{ij} - \frac{c_{..}}{N^2}\right) = 0$.

3. Lien du critère E et du critère $B(k)$

3.1. Présentation de l'indice B_k

Récemment E.B. Fowlkes et C.L. Mallows 1983 [4] ont défini une mesure de similarité entre deux partitions issues de hiérarchies (mais il est également applicable pour les partitions directes) notée B_k où k désigne les niveaux de la hiérarchie. L'expression en notations par paires du critère B_k s'écrit :

$$B_k = \frac{\sum_{ij} c_{ij} y_{ij}}{\sqrt{c_{..} y_{..}}}.$$

B_k s'interprète comme une corrélation entre les vecteurs de composantes c_{ij} et y_{ij}, ou encore comme une pondération des associations positives par la moyenne géométrique des tableaux C et Y.

Avec les mêmes notations que celles considérées au départ, on obtient, pour k fixé :

$$B_k = \frac{\sum_{u,v} n_{uv}^2}{\sqrt{\left(\sum_v n_{.v}^2 \sum_u n_{u.}^2\right)}}.$$

Dans son commentaire sur l'indice B_k, Wallace [4] a défini deux indices W_1 et W_2 tel que :

$$W_1 = \frac{\sum_{u,v} n_{uv}^2}{\sum_v n_{.v}^2}; \quad W_2 = \frac{\sum_{u,v} n_{uv}^2}{\sum_u n_{u.}^2}.$$

L'indice B_k, comme on le remarque, s'interprète comme la moyenne géométrique de ces deux indices W_1 et W_2. Aucun de ces deux indices ne s'annule en cas d'indépendance statistique; de ce fait, nous allons utiliser la normalisation présentée précédemment sur W_1 et W_2. On aura alors en appelant $\widetilde{W}_1$ et $\widetilde{W}_2$ les indices normalisés ainsi obtenus.

$$\widetilde{W}_2 = \frac{\sum_{u,v} n_{uv}^2 - \frac{\sum_u n_{u.}^2 \sum_v n_{.v}^2}{N^2}}{\sum_u n_{u.}^2\left(1 - \frac{\sum_v n_{.v}^2}{N^2}\right)} \qquad \widetilde{W}_1 = \frac{\sum_{u,v} n_{uv}^2 - \frac{\sum_u n_{u.}^2 \sum_v n_{.v}^2}{N^2}}{\sum_v n_{.v}^2\left(1 - \frac{\sum_u n_{u.}^2}{N^2}\right)}.$$

3.2. Proposition

$$\frac{E}{\sqrt{E_1 \times E_2}} = \sqrt{\widetilde{W}_1 \times \widetilde{W}_2}$$

En effet

$$\sum_u n_{u.}^2 \left(1 - \frac{\sum_v n_{.v}^2}{N^2}\right) \sum_v n_{.v}^2 \left(1 - \frac{\sum_u n_{u.}^2}{N^2}\right) =$$

$$= \sum_u n_{u.}^2 \left(1 - \frac{\sum_u n_{u.}^2}{N^2}\right) \sum_v n_{.v}^2 \left(1 - \frac{\sum_v n_{.v}^2}{N^2}\right);$$

on peut donc interpréter la corrélation entre les vecteurs $\left(c_{ij} - \frac{c_{..}}{N^2}\right)$ et $\left(y_{ij} - \frac{y_{..}}{N^2}\right)$ comme étant la moyenne géométrique des indices de Wallace corrigés, en cas d'association complète $\frac{E}{E_{\max}} = \sqrt{\widetilde{W}_1 \times \widetilde{W}_2}$ avec $E_{\max} = \sqrt{E_1 E_2}$.

4. Le critère E dans un contexte prédictif

Présentation

Considérons deux partitions C et Y et les tableaux associés de comparaison par paires que l'on notera également C et Y.

On désignera par R_{CY} (resp. R_{CY}) le coefficient de régression de Y (resp. C) par rapport à C (resp. Y) quand on veut expliquer Y par C (resp. C par Y).

Compte tenu de ce que $y_{ij}^2 = y_{ij}$ et $c_{ij}^2 = c_{ij}$, les formules classiques donnant R_{CY} et R_{CY} se simplifient, et l'on obtient (puisqu'on a $n = N^2$ valeurs pour les y_{ij} et c_{ij}) :

$$R_{YC} = \frac{N^2 \sum_{ij} c_{ij} y_{ij} - \sum_{ij} c_{ij} \sum_{ij} y_{ij}}{\sum_{ij} y_{ij} \left(N^2 - \sum_{ij} y_{ij}\right)}, \quad R_{CY} = \frac{N^2 \sum_{ij} c_{ij} y_{ij} - \sum_{ij} c_{ij} \sum_{ij} y_{ij}}{\sum_{ij} c_{ij} \left(N^2 - \sum_{ij} c_{ij}\right)}$$

D'où contingentiellement R_{CY} et R_{YC} à partir des formules de transfert (paires-contingence) s'écrivent :

$$R_{CY} = \frac{\sum_{u,v} n_{uv}^2 - \frac{\sum_u n_{u.}^2 \sum_v n_{.v}^2}{N^2}}{\sum_u n_{u.}^2 \left(1 - \frac{\sum_u n_{u.}^2}{N^2}\right)}; \quad R_{YC} = \frac{\sum_{u,v} n_{uv}^2 - \frac{\sum_u n_{u.}^2 \sum_v n_{.v}^2}{N^2}}{\sum_v n_{.v}^2 \left(1 - \frac{\sum_v n_{.v}^2}{N^2}\right)} \tag{5}$$

Remarque :

Dans sa formulation en paires le critère R_{CY} est linéaire en y_{ij}, et sera un critère utile dans les problèmes d'association maximale que nous étudierons ultérieurement

et qui seront abordés dans [9], [10], en effet :

$$R_{CY} = \frac{1}{K}\sum_{ij}\left(c_{ij} - \frac{c_{..}}{N^2}\right)y_{ij} \quad \text{où} \quad K = c_{..}\left(1 - \frac{c_{..}}{N^2}\right).$$

Il est important de constater que si C est donné, on a alors la solution des moindres carrés qui est donnée par :

$$\hat{Y} = \hat{A}C + \hat{B} = R_{CY}C + \hat{B},$$

avec

$$\hat{A} = E/E_1 \quad \text{et} \quad \hat{B} = \left(\sum_{ij} c_{ij}y_{ij} - c_{..}y_{..}/N^2\right)/(c_{..} - N^2) + y_{..}/N^2.$$

En d'autres termes le critère E, dont l'interprétation fondamentale en paires est liée à sa signification de critère d'adéquation pondérée de deux préordonnances voir [2] S. Chah, joue un rôle important comme critère d'association. Son rôle, peu connu des statisticiens est essentiellement fondé sur sa position intermédiaire entre différents critères plus connus (Rand, χ^2) et d'agent de liaison interprétative entre leurs comportements et significations *a priori* totalement différentes.

5. Le critère I, l'écart à l'indétermination

5.1. Présentation

En utilisant les notations associées à un tableau de contingence, l'indétermination se traduit par la configuration suivante :

$$n_{uv} = \frac{n_{u.}}{q} + \frac{n_{.v}}{p} - \frac{N}{pq} \quad \forall(u,v).$$

La plupart des critères d'association statistique sont basés sur la notion d'écart de chaque case n_{uv} du tableau d'indépendance, c'est-à-dire au cas où $n_{uv} = n_{u.}n_{.v}/N$; ici on s'intéressera plutôt à l'écart de chaque case n_{uv} à l'indétermination. Le critère que l'on étudiera s'insèrera dans la liste des différents critères d'associations. Le critère a été étudié par F. Marcotorchino dans [8], il s'écrit contingentiellement :

$$I = \sum_{u,v} n_{uv}^2 - \frac{\sum_u n_{u.}^2}{q} - \frac{\sum_v n_{.v}^2}{p} + \frac{N^2}{pq} = \sum_{u,v}\left[n_{uv} - \left(\frac{n_{u.}}{q} + \frac{n_{.v}}{p} - \frac{N}{pq}\right)\right]^2$$

La deuxième formulation montre bien qu'il sagit de l'écart carré à la sructure d'indétermination.

5.2. Propriétés

- Le critère I est nul en cas d'indétermination (par construction).

- I est maximal en cas d'association complète (i.e.) : $p = q$, $n_{u,v} = n_{u.} = n_{.v}$ si $u = v$, $n_{u,v}$ si $u \neq v$.

• La valeur maximale de l'indice est alors donnée par l'une ou l'autre des 2 quantités suivantes qui sont égales :

$$I_1 = \left(1 - \frac{2}{p}\right) \sum_u n_{u.}^2 + \frac{N^2}{p^2}$$

$$I_2 = \left(1 - \frac{2}{q}\right) \sum_v n_{.v}^2 + \frac{N^2}{q^2}$$

Si on est pas dans le cas de l'association complète $I_1 \neq I_2$.

• Le critère a également un rapport très étroit avec la décomposition en variance du tableau de contingence croissant C et Y. En effet, dans le cas d'une classification croisée sans répétition, et en identifiant la réalisation habituelle X_{ij} à la valeur n_{uv}, les formules de décomposition de la variance totale nous donnent voir par exemple [8] :

$$S = \sum_{u,v} \left(n_{uv} - \frac{N}{p \times q}\right)^2, \quad S_1 = \frac{1}{q} \sum_{u,v} \left(n_{u.} - \frac{N}{p}\right)^2, \quad S_2 = \frac{1}{p} \sum_{u,v} \left(n_{.v} - \frac{N}{q}\right)^2.$$

En se servant des expressions précédentes on retrouve la formule classique de la décomposition des variances en analyse de la variance à savoir :

$$S = S_1 + S_2 + \sum_{u,v} n_{uv}^2 - \frac{\sum_u n_{u.}^2}{q} - \frac{\sum_v n_{.v}^2}{p} + \frac{N^2}{pq} = S_1 + S_2 + I.$$

Le critère I joue le rôle de la variance résiduelle.

S la variance totale du tableau.
$S_1 =$ variance du tableau par rapport à C,
$S_2 =$ variance du tableau par rapport à Y.

I mesure l'interaction qui existe entre C et Y, plus cette interaction est forte plus Y sera en association avec C.

5.3. Écriture en paires du critère I

En paires les critères I, I_1 et I_2 s'écrivent :

$$I = \sum_{ij} \left(c_{ij} - \frac{1}{p}\right)\left(y_{ij} - \frac{1}{q}\right), \quad I_1 = \sum_{ij} \left(c_{ij} - \frac{1}{p}\right)^2, \quad I_2 = \sum_{ij} \left(y_{ij} - \frac{1}{q}\right)^2.$$

Dans cette nouvelle formulation I s'interprète comme une covariance (non centrée et au facteur $1/N^2$ près) entre les vecteurs de composantes $c_{ij} - \frac{1}{p}$ et $y_{ij} - \frac{1}{q}$.

Remarque

Si $c_{..} = \frac{N^2}{p}$ alors $E = I$.

$$E - I = \sum_{ij} \left(c_{ij} - \frac{c_{..}}{N^2}\right)\left(y_{ij} - \frac{1}{q}\right) - \sum_{ij} \left(c_{ij} - \frac{1}{p}\right)\left(y_{ij} - \frac{1}{q}\right)$$

soit

$$E - I = \left(1/p - \frac{c_{..}}{N^2}\right) \sum_{ij} \left(y_{ij} - \frac{1}{q}\right)$$

donc

$$E - I = K \sum_{ij} \left(y_{ij} - \frac{1}{q}\right)$$

où

$$K = \left(1/p - \frac{c_{..}}{N^2}\right)$$

si $\sum_{ij} \left(y_{ij} - \frac{1}{q}\right)$ est constant, alors

$$\max_Y E = \max_Y I.$$

6. Le critère J de Végelius et Janson

6.1. Présentation

Récemment un nouveau coefficient de corrélation pour variables qualitatives a été introduit par S. Janson et J. Végelius [7]. L'indice J est un cas spécial du coefficient généralisé de Kendall. Janson et Végelius considèrent l'indice j comme une mesure d'accord.

Contingentiellement le critère J s'écrit : si $p \geq 2$ et $q \geq 2$

$$J = \frac{pq \sum_{u,v} n_{uv}^2 - p \sum_u n_{u.}^2 - q \sum_v n_{.v}^2 + N^2}{\sqrt{\left[p(p-2) \sum_u n_{u.}^2 + N^2\right] \left[q(q-2) \sum_v n_{.v}^2 + N^2\right]}}. \tag{6}$$

D'autre part on pourrait transformer directement la formule (6) en utilisant les formules de passage (contingence-paires) on obtient alors :

$$J = \frac{\sum_{ij} \left(c_{ij} - \frac{1}{p}\right)\left(y_{ij} - \frac{1}{q}\right)}{\sqrt{\left[\sum_{ij} \left(c_{ij} - \frac{1}{p}\right)^2 \left(y_{ij} - \frac{1}{q}\right)^2\right]}} \tag{7}$$

J est donc un cosinus entre les vecteurs de composantes $c_{ij} - 1/p$ et $y_{ij} - 1/q$.

6.2. Propriétés du critère J

- J est nul en cas d'indétermination.
- J est symétrique.
- J vérifie la propriété de la transitivité de l'association parfaite : $J(A,B) = 1$ et $J(A,C) = 1$ alors $J(B,C) = 1$; $J(A,A) = 1$.
- La matrice de corrélation basée sur l'indice J est définie positive.

Pour plus de détails sur l'indice J, voir [7].

7. Lien entre le critère I et l'indice J

On remarque que le critère J n'est qu'une normalisation du critère I, en effet d'après la formule (7) on a $J = \frac{I}{\sqrt{I_1 I_2}}$. considérons les deux critères non symétriques suivants :

$$I_{N1} = \frac{I}{I_1}, \qquad I_{N2} = \frac{I}{I_2}.$$

L'introduction de I_{N1} et I_{N2} a un intérêt dans le cas où une partition est constante et si l'on veut prédire l'autre avec I_{N1} ou I_{N2} maximal, en effet :

$$I_{N1} = \frac{1}{K} \sum_{ij} \left(c_{ij} - \frac{1}{p}\right)\left(y_{ij} - \frac{1}{q}\right)$$

$$I_{N2} = \frac{1}{L} \sum_{ij} \left(c_{ij} - \frac{1}{p}\right)\left(y_{ij} - \frac{1}{q}\right)$$

$$K = \sum_{ij} \left(c_{ij} - \frac{1}{p}\right)^2$$

$$L = \sum_{ij} \left(y_{ij} - \frac{1}{q}\right)^2.$$

L'avantage de I_{N1} par rapport à J réside dans le fait que le dénominateur ne dépend pas de Y. Ceci peut être intéressant lorsque on utilise les concepts d'association maximale avec ces critères, voir [8].

Le critère J, normalisation du critère I, est peu utilisé, sauf par l'école scandinave de satistiques; néanmoins certains travaux récents, portant sur l'étude comparative du coefficient J et du coefficient χ^2, montre que le coefficient J a asymptotiquement un comportant voisin du χ^2, voir [7]. Une autre étude comparative a été faite par R. Popping dans [10]. La comparaison faite à l'aide de simulation de Monte-Carlo, sur les indices J, D_2 (Nominal Scale Agreement R. Popping), montre que $J(I)$ et D_2 ont un comportement identique, D_2 étant défini par :

$$D_2 = (d_0 - d_{\exp})/(d_{\max} - d_{\exp})$$

où d_0 sont les accords positifs, $d_{\max}$ la valeur maximale de d_0 et $d_{\exp}$ la valeur estimée de d_0 sous la condition de l'indépendance.

8. Le critère de Belson

Contingentiellement le critère de Belson s'écrit :

$$B(C,Y) = \sum_{u,v} \left(n_{uv} - \frac{n_{u.} n_{.v}}{N}\right)^2.$$

Comme on le remarque $B(C,Y)$ mesure l'éloignement de chaque case du tableau n_{uv} à la structure d'indépendance statistique, $n_{uv} = n_{u.}n_{.v}/N$.

Propriétés

- Le critère B est nul en cas d'indépendance statistique (par construction).
- $B(C,Y) \geq 0$.
- B est maximal en cas d'association complète (i.e.) : $p = q, n_{u,v} = n_{u.} = n_{.v}$ si $u = v, n_{u,v} = 0$ si $u \neq v$.

La valeur maximale de l'indice est alors donnée par l'une ou l'autre des 2 quantités suivantes qui sont égales :

$$B_1 = \sum_u n_{u.}^2 - 2\sum_u \frac{n_{u.}^3}{N} + \left(\sum_u \frac{n_{u.}^2}{N}\right)^2$$

$$B_2 = \sum_v n_{.v}^2 - 2\sum_v \frac{n_{.v}^3}{N} + \left(\sum_u \frac{n_{.v}^2}{N}\right)^2.$$

Ces valeurs ne sont égales que dans le cas de l'association complète.

Si l'on pose $B_{\max}$ la valeur du critère en cas d'association complète, $B_{\max} = \sqrt{B_1 \times B_2}$.

En utilisant les formules de passage contingence paires, avec la même désignation, B s'écrit :

$$B(C,Y) = \sum_{ij}\left(c_{ij} - \frac{c_{i.} + c_{.j}}{N} + \frac{c_{..}}{N^2}\right)y_{ij}.$$

Si l'on pose $t_{ij} = \left(c_{ij} - \frac{c_{i.} + c_{.j}}{N} + \frac{c_{..}}{N^2}\right)$, alors T est la décomposition de TORGERSON du tableau C.

Remarque 1

B est une covariance entre les vecteurs de composantes t_{ij} et y_{ij}.

Remarque 2

Dans [14] G.Saporta a défini un indice angulaire entre opérateurs associés à des variables qualitatives. Cet indice noté S s'écrit :

$$S = \frac{\text{trace}\,(V_{12}V_{21})}{\sqrt{\text{trace}\,V_{11}^2\ \text{trace}\,V_{22}^2}}$$

V_{12} étant la matrice de terme général $\frac{1}{N}\left(n_{uv} - \frac{n_{u.}n_{.v}}{N}\right)$ et V_{21} sa matrice transposée.

V_{11} est la matrice de terme diagonaux $\frac{n_{u.}}{N}\left(1 - \frac{n_{u.}}{N}\right)$ et de termes non diagonaux $-\frac{n_{u.}n_{u'.}}{N^2}$.

V_{22} étant la matrice analogue à V_{11} portant sur $n_{.v}$.

L'indice angulaire associé développé en notation contingentielle devient :

$$S = \frac{1}{N^2}\sum_{u,v}\left(n_{uv} - \frac{n_{u.}n_{.v}}{N}\right)^2 \times \frac{1}{\sqrt{\sum_u\left(\frac{n_{u.}}{N}\left(1-\frac{n_{u.}}{N}\right)\right)^2 + \left(\sum_{\substack{u,u' \\ u\neq u'}}\frac{n_{u.}}{N}\frac{n_{u'.}}{N}\right)^2}} \times \frac{1}{\sqrt{\sum_v\left(\frac{n_{.v}}{N}\left(1-\frac{n_{.v}}{N}\right)\right)^2 + \left(\sum_{\substack{v,v' \\ v\neq v'}}\frac{n_{.v}}{N}\frac{n_{.v'}}{N}\right)^2}}.$$

Après simplification, on peut constater que :

$$S = \frac{\sum_{u,v}\left(n_{uv} - \frac{n_{u.}n_{.v}}{N}\right)^2}{\sqrt{\left(\sum_u n_{u.}^2 - 2\sum_u \frac{n_{u.}^3}{N} + \left(\sum_u \frac{n_{u.}^2}{N}\right)^2\right)\left(\sum_v n_{.v}^2 - 2\sum_v \frac{n_{.v}^3}{N} + \left(\sum_v \frac{n_{.v}^2}{N}\right)^2\right)}}.$$

C'est-à-dire $B/\sqrt{B_1 \times B_2} = S$ c'est-à-dire la moyenne géométrique des deux critères non symétriques $\frac{B}{B_1}$ et $\frac{B}{B_2}$ est égale à l'indice angulaire, d'où le lien entre B et S.

9. Conclusion comparative

Après la présentation des indices E, MOREY AGRESTI (Ω cf. ci-dessous), B_k, I, J, S, B, il apparaît que :

$\Omega = E/\sqrt{E_1 \times E_2}, J = I/\sqrt{I_1 \times I_2}, S = B/\sqrt{B_1 \times B_2}$.

En d'autres termes les indices Ω, J et S sont des normalisations par moyennes géométriques des maximas possibles des indices sources (E, I et B). Seule donc l'étude des indices de base E, I, B est nécessaire à la formation des indices dérivés.

10. Normalisation de quelques critères

Dans ce paragraphe, nous allons essayer de montrer comment le principe de normalisation défini dans l'introduction, va nous permettre une homogénéisation de l'interprétation et du comportement de certains des critères d'associations étudiés précédemment.

Appliquons en effet la normalisation définie précédemment aux critères de Rand (R), Fowlkes et Mallows (B_k), Écart à la moyenne (E).

10.1. Le critère de Rand

Sur un tableau de contingence T, R s'écrit dans sa forme généralisée donnée dans [8].

$$R = \frac{2\sum_{u,v} n_{uv}^2 - \left(\sum_u n_{u.}^2 + \sum_v n_{.v}^2\right)}{n^2} + 1.$$

En utilisant le procédé de normalisation, on obtient :

$$R_N = \frac{\sum_{u,v} n_{uv}^2 - \dfrac{\sum_u n_{u.}^2 \sum_v n_{.v}^2}{N^2}}{\dfrac{\sum_u n_{u.}^2 + \sum_v n_{.v}^2}{2} - \dfrac{\sum_u n_{u.}^2 \sum_v n_{.v}^2}{N^2}} = \frac{R - \widehat{R}}{1 - \widehat{R}} = \Omega \qquad (8)$$

où Ω est le critère de Morey et Agresti.

Mais ce qui est intéressant ici c'est le fait de constater que seule la normalisation du terme $A_1 = \sum_{u,v} n_{uv}^2 = \sum_{ij} c_{ij} y_{ij}$ intervient dans (8). En effet R_N s'interprète comme une normalisation du critère A_1 en prenant comme borne la moyenne arithmétique des tableaux de comparaison par paires C et Y; en effet si l'on pose $f_R(n_{u.}, n_{.v})$ la borne de A_1 on a :

$$f_R(n_{u.}, n_{.v}) = \frac{\sum_v n_{u.}^2 + \sum_u n_{.v}^2}{2} = \frac{(c_{..} + y_{..})}{2}.$$

Si on utilise l'inégalité de Cauchy-Schwartz on aura :

$$\sum_{u,v} n_{uv}^2 = \sum_{ij} c_{ij} y_{ij} \leq \sqrt{\sum_{ij} c_{ij} \sum_{ij} y_{ij}} \leq \frac{\sum_{ij} c_{ij} + \sum_{ij} y_{ij}}{2} = \frac{(c_{..} + y_{..})}{2}.$$

Remarque

$$\frac{R - \widehat{R}}{1 - \widehat{R}} = \frac{A_1 - \widehat{A}_1}{A_{1\max} - \widehat{A}_1} \quad \text{où} \quad A_{1\max} = (c_{..} + y_{..})/2.$$

La normalisation est une opération transparente sur R en ce sens qu'elle porte que sur A_1.

10.2. Le critère de Fowlkes et Mallows

Nous avons vu précédemment que B_k généralisé est donné par :

$$B_k = \frac{\sum_{u,v} n_{uv}^2}{\sqrt{\left(\sum_v n_{.v}^2 \sum_u n_{u.}^2\right)}} \quad 0 \leq B_k \leq 1.$$

En utilisant le procédé de normalisation, on obtient :

$$B_{kN} = \frac{\sum_{u,v} n_{uv}^2 - \dfrac{\sum_u n_{u.}^2 \sum_v n_{.v}^2}{N^2}}{\sqrt{\sum_v n_{.v}^2 \sum_u n_{u.}^2} - \dfrac{\sum_v n_{u.}^2 \sum_u n_{.v}^2}{N^2}} = \frac{B_k - \widehat{B}_k}{1 - \widehat{B}_k} \qquad (9)$$

B_{kN}, comme on le remarque, est une normalisation du critère A_1, en prenant comme borne la moyenne géométrique de tableaux de comparaison par paires C et Y; $f_B(n_{u.}, n_{.v} = \sqrt{\sum_u n_{.v}^2 \sum_v n_{u.}^2} = \sqrt{c_{..} \times y_{..}}$.

10.3. L'écart à la moyenne normalisé

Considérons les deux indices de régressions non symétriques définis précédemment (cf. formule (5)). Ces indices qui sont déjà normalisés peuvent encore s'écrire :

$$R_{CY} = \frac{\sum_{u,v} n_{uv}^2 - \frac{\sum_u n_{u.}^2 \sum_v n_{.v}^2}{N^2}}{\sum_u n_{u.}^2 \left(1 + \frac{\sum_v n_{.v}^2 - \sum_u n_{u.}^2}{N^2}\right) - \frac{\sum_u n_{u.}^2 \sum_v n_{.v}^2}{N^2}} \tag{10}$$

$$R_{CY} = \frac{\sum_{u,v} n_{uv}^2 - \frac{\sum_u n_{u.}^2 \sum_v n_{.v}^2}{N^2}}{\sum_v n_{.v}^2 \left(1 + \frac{\sum_u n_{u.}^2 - \sum_v n_{.v}^2}{N^2}\right) - \frac{\sum_u n_{u.}^2 \sum_v n_{.v}^2}{N^2}} \tag{11}$$

R_{CY} ou R_{YC} s'interprètent comme des normalisations du critère A_1, avec comme borne les quantités :

$$f_R(n_{u.}, n_{.v}) = \sum_u n_{u.}^2 \left(1 + \frac{\sum_v n_{.v}^2 - \sum_u n_{u.}^2}{N^2}\right)$$

ou encore

$$f_R(n_{u.}, n_{.v}) = \sum_v n_{.v}^2 \left(1 + \frac{\sum_u n_{u.}^2 - \sum_v n_{.v}^2}{N^2}\right).$$

Remarque 1

Les trois critères étudiés ici se comportent comme si l'on n'avait à considérer que A_1. Ils ne diffèrent que dans le choix de la borne que l'on va affecter à A_1.

Remarque 2

Dans les trois normalisations précédentes le critère d'écart à la moyenne apparaît au numérateur.

$$R_N = \frac{E}{D_1}, \quad B_{kN} = \frac{E}{D_2}, \quad R_{CY} = \frac{E}{D_3},$$

D_1, D_2, D_3 sont les dénominateurs des expressions R_N, B_{kN}, R_{CY}.

Remarque 3

Considérons dans le même ordre d'idée une pondération du critère A_1 de la façon suivante :

$$A_1' = \sum_{u,v} \frac{N}{n_{u.}} n_{uv}^2. \tag{12}$$

Sachant que pour (u, v), $n_{uv} n_{u.} \geq n_{uv}^2$ alors $A_1' \leq N^2$.

Si l'on prend N^2 comme borne de A_1' et en utilisant la normalisation définie précédemment on obtient :

$$A_{1N}' = \frac{\sum_{u,v} \frac{N}{n_{u.}} n_{uv}^2 - \sum_v n_{.v}^2}{N^2 - \sum_v n_{.v}^2} = \frac{A_1' - \widehat{A}_1'}{A_{1\max}' - \widehat{A}_1'} \quad \text{où} \quad \widehat{A}_1' = \sum_v n_{.v}^2. \tag{13}$$

On retrouve le critère de prédiction proportionnelle τ_b de Goodman et Kruskal. En conclusion le critère τ_b s'interprète comme une normalisation de A_1', pondération des accords positifs.

Après avoir utilisé une normalisation par rapport à la structure d'indépendance, on pourra appliquer ce procédé pour centrer les critères par rapport à la structure d'indétermination. Soit H un indice, on notera $\widetilde{H}$ l'indice normalisé, $\widehat{H}$ la valeur du critère en cas d'indétermination :

$$n_{uv} = \frac{n_{u.}}{q} + \frac{n_{.v}}{p} - \frac{N}{pq}$$

alors $\widetilde{H}$ est défini par : $\widetilde{H} = \dfrac{H - \widehat{H}}{H_{\max} - \widehat{H}}$. Il s'annule en cas d'indétermination,

$$\widetilde{H} = 0 \quad \text{si } H = \widehat{H}$$

$$\widetilde{H} = 1 \quad \text{si } H = H_{\max}.$$

$\widetilde{H}$ est une normalisation de H. Les indices étudiés antérieurement s'écrivent $\sum_{u,v} n_{uv}^2 + g(n_{u.}, n_{.v})$ de ce fait $\widetilde{H}$ s'écrit :

$$\widetilde{H} = \frac{\sum_{u,v} n_{uv}^2 - \frac{\sum_u n_{u.}^2}{q} - \frac{\sum_v n_{.v}^2}{p} + \frac{N^2}{pq}}{H_{\max} - \frac{\sum_u n_{u.}^2}{q} - \frac{\sum_v n_{.v}^2}{p} + \frac{N^2}{pq}}. \tag{14}$$

On retrouve ainsi des normalisations du critère des associations positives A_1 par rapport à l'indétermination. À titre d'exemple on obtient pour le critère de Rand :

$$\widetilde{R} = \frac{\sum_{u,v} n_{uv}^2 - \frac{\sum_u n_{u.}^2}{q} - \frac{\sum_v n_{.v}^2}{p} + \frac{N^2}{pq}}{\frac{\sum_u n_{u.}^2 + \sum_v n_{.v}^2}{2} - \frac{\sum_u n_{u.}^2}{q} - \frac{\sum_v n_{.v}^2}{p} + \frac{N^2}{pq}}. \tag{15}$$

Remarque 4

La valeur $\widetilde{H}$ est la valeur minimale du critère A_1 lorsque les marges sont fixées. D'où :

$$\widetilde{R} = \frac{\sum_{u,v} n_{uv}^2 - (A_1)_{\min}}{\frac{\sum_u n_{u.}^2 + \sum_v n_{.v}^2}{2} - (A_1)_{\min}}. \tag{16}$$

11. Récapitulatif de quelques critères d'association

a	b	Q	identification du critère
0	0	$\sqrt{c_{..}y_{..}}$	Fowlkes et Mallows B_k
$1/2$	$1/2$	$N^2/4$	l'indice de L. Hubert Γ
$1/p$	$1/q$	1	l'écart à l'indétermination I
$1/p$	$1/q$	$\sqrt{X \times Y}$ $X = \sum_{ij}\left(c_{ij} - \frac{1}{p}\right)^2$ $Y = \sum_{ij}\left(y_{ij} - \frac{1}{q}\right)^2$	le critère de Vegelius J
1	1	$N^2/2$	le critère de l'information mutuelle quadratique I^2 [4]
$c_{..}/N^2$	$b \in R$	1	écart à la moyenne E
$c_{..}/N^2$	$c_{..}/N^2$	$c_{..}(1 - c_{..}/N^2)$	coefficient de régression R_{CY}
$(c_{i.} + c_{.j})/2$	0	1	le critère de dérivé de Jordan [8] J_0 ou Con-Dis (L.Hubert) mesure d'association entre triplets de deux partitions sur le même ensemble d'éléments.
$(c_{i.} + c_{.j})/2 - c_{..}/N^2$	0	1	le critère de Belson généralisé B [8]
$1/N$	0	1 ou $\sum_{ij} \bar{y}_{ij}$	si l'on pondère les c_{ij} par $c_{i.} + c_{.j}/2$ on retrouve le critère de Light [8] ou τ_b de Goodman-Kruskal

Cette récapitulation se fera au travers des schémas normatifs suivants :

$$\Lambda_1 = \frac{\sum_{i,j}(c_{ij} - a)(y_{ij} - b)}{Q} \qquad \Lambda_2 = \frac{\sum_{i,j}(s_{ij} - a)(y_{ij} - b)}{Q} \tag{17}$$

s_{ij} est une pondération de éléments c_{ij} du tableau de comparaison par paires. Seules les valeurs a, b et Q nous intéresseront et changeront. Nous avons montré ici en utilisant le principe de linéarisation des indices introduit par F. Marcotorchino dans [8] et les résultats donnés dans cet article que les indices précédemment étudiés plus quelques autres s'inscrivent dans la forme donnée en (17).

12. Conclusion

Tout au long de ce travail, nous avons eu pour objectif de démontrer que des liens existaient entre certains critères d'association. Nous avons vu en particulier les liens entre le critère de l'écart à l'indétermination et le critère de Végélius, le critère de Belson et le critère proposé par G. Saporta, l'écart à la moyenne et le critère de Fowlkes et Mallows etc ...

Pour ce faire, nous avons fait appel aux procédés de linéarisation et de normalisation. Ils ont permis de mettre en évidence les parentés mathématiques des différents critères. Nous avons constaté, à travers cette étude que le critère des accords positifs jouait un rôle prépondérant. Par le biais de $A_1 = \sum_{ij} c_{ij} y_{ij}$ on peut aboutir à la plupart des indices d'associations en normalisant A_1.

Références

[1] A. AGRESTI et L. MOREY (1984). — An adjustement of the Rand statistic for chance agreement, *Educational and Psychological Measurement*, 44, 33–37.

[2] S. CHAH (1983). — Optimisation linéaire en classification automatique, thèse de troisième cycle, Paris VI.

[3] E.B. FAWLKES MALLOWS (1983). — a Method for Comparing two Hierarchical Clusterings, *JASA*, Vol.78, 553–584.

[4] M.A. GIL, R. PEREZ et I. MARTINEZ (1986). — The mutuel information estimation in the sampling with replacement. *R.A.I.R.O*, n° 20, 3, 257–268.

[5] P. GREEN, V.R. RAO (1969). — Note Proximity Measures and Cluster Analysis, Journal of Marketing Research, Vol. 6, p. 359–364.

[6] L. HUBERT (1977). — Nominal Scale Response Agrement as a Generalised Correlation, *B.J. Mat. Sta. Psy*, 30, p. 98–103.

[7] S. JANSON et J. VEGELIUS (1982). — Criteria for symetric measures of association for Nominal Data, *Quality and Quantity*, Vol. 16, p. 243–250.

[8] F. MARCOTORCHINO (1984). — Utilisation des Comparaisons par Paires en Statistique des Contingences, *Étude du Centre Scientifique IBM- France* , n° F-071.

[9] F. MARCOTORCHINO (1986). — Maximal Association Theory as a Tool of Research, in W. Gaul and M. Schader (editors) Proceeding of the 9th Annual meeting of the classification society (F.R.G), North- Holland.

[10] F. MARCOTORCHINO (1986). — Cross Association Measures and Paired Comparaisons, *Proceedings of COMPSTAT Symposium*, p. 286. Physika Verlag.

[11] B.G. MIRKIN (1970). — Measurement of the distance between distincts partitions of a finite set of objets, *Automation and Remote Control*, Vol. 31, p. 786–792.

[12] R. POPPING (1984). — Traces of Agreement : On some Agreement Indices for Open ended Questions, *Quality and Quantity*, 18, p. 147–158.

[13] W. M. RAND (1971). — Objective Criteria for the Evaluation of Clustering Method, *JASA*, Vol. 66, p. 846–850.

[14] G. SAPORTA (1975). — Liaisons entre plusieurs Ensembles de Variables et Codages de Données Qualitatives, Thèse troisième cycle Paris VI.