MATHÉMATIQUES ET SCIENCES HUMAINES

ISRAËL-CÉSAR LERMAN

Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles. II

Mathématiques et sciences humaines, tome 119 (1992), p. 75-100

<http://www.numdam.org/item?id=MSH_1992__119__75_0>

Math. Inf. Sci. hum. (29e année, n°119, *1992*, p.75-100)

CONCEPTION ET ANALYSE DE LA FORME LIMITE D'UNE FAMILLE DE COEFFICIENTS STATISTIQUES D'ASSOCIATION ENTRE VARIABLES RELATIONNELLES . II

Israël-César LERMAN[1]

RÉSUMÉ — *Cette étude offre une large vision de synthèse prospective ; mais aussi, des résultats techniques précis sur une famille très générale que nous avons élaborée de coefficients d'association entre variables descriptives relationnelles à partir de leur observation empirique sur un ensemble O d'objets élémentaires. Un même coefficient est obtenu à partir d'une forme de normalisation statistique par rapport à une hypothèse d'absence de liaison, d'un indice brut d'association. Ce dernier suppose une représentation de type ensembliste des deux variables relationnelles à comparer. Le cas où les deux variables sont unaires introduit et pose clairement le problème. Nous étudions particulièrement le cas où les deux relations induites par les deux variables sont binaires. Ce cas est d'une extrême utilité en analyse des données qualitatives. La normalisation suppose le centrage et la réduction par l'écart type de l'indice brut aléatoire. C'est une expression particulière de la variance de ce dernier qui permet de mettre en évidence la forme limite du coefficient d'association dans des conditions qu'on appréhende clairement. On considère avec soin les cas très importants de la comparaison de deux variables qualitatives nominales ou ordinales. L'expression limite permet de se rendre compte d'un point de vue purement formel de la nature de la normalisation ainsi effectuée. Nous abordons ensuite un cas assez général de comparaison de deux relations q-aires pour lequel l'essentiel des calculs est fourni. Enfin, nous exprimons les recherches actuelles et développements futurs, en situant la place de ce travail dans l'aspect "classification hiérarchique" de notre approche en analyse des données.*

SUMMARY — Elaboration and analysis of the limit form of a family of statistical association coefficients between relational variables.

This study gives a large synthesis view and prospective on a very general family of association coefficients between descriptive relational variables, that we have elaborated. On the other hand, very accurate technical results are provided. We assume the empirical observation of the descriptive variables on a set O of elementary objects. A given coefficient is obtained by a statistical normalization of a raw association index with respect to a hypothesis of no relation (or independence). The raw index s is conceived from a set theoretic representation of the two relational variables to be compared. The case where the two variables associated are unary, provides a clear setting up of the comparison problem. We particularly analyze the case where the two relations on O, induced by the two descriptive variables to be compared, are binary. The latter case is extremely useful in qualitative data analysis. The normalization of the raw index s takes into account the distribution of the random raw index S under an independence hypothesis. The reduction of the "centred" index [s-E(S)] where E denotes the mathematical expectation] is done with the standard deviation var(S). It is a specific expression of the variance var(S), which enables to set up the limiting from of an association coefficient, under natural asymptotic conditions. Then, we carefully study the very important cases where the descriptive variables are nominal or ordinal qualitative variables. The limit expression permits to realize the nature of the normalization, from a purely formal point of view. Next, we take up the study of the general case of the comparison of two q-ary relations. Accurate results are given in the latter context. Finally, we express our current research and their future development ; more particularly by situating the place of this work in our approach of data analysis by means of hierarchical classification.

[1] IRISA, Campus de Beaulieu, 35042 RENNES Cedex.

4. FORME LIMITE DU COEFFICIENT D'ASSOCIATION

4.1. Les deux expressions de base de la variance de l'indice brut aléatoire S_1

L'expression de S_1 est donnée en (16) du paragraphe 3 ci-dessus et nous avons déjà mentionné que sa distribution est la même que celle des deux indices aléatoires duaux

$$s(\sigma,I) = \Sigma\left\{x_{\sigma(i)\,\sigma(j)}y_{ij} / (i,j) \in I^{[2]}\right\} \tag{1}$$

et

$$s(I_d,\tau) = \Sigma\left\{x_{ij}\,y_{\tau(i)\tau(j)} / (i,j) \in I^{[2]}\right\} \tag{2}$$

où $I^{[2]}$ a été spécifié en (7) du paragraphe 3 ci-dessus et où σ (resp.τ) est un élément aléatoire dans l'ensemble G*n* muni d'une probabilité uniforme, de toutes les permutations sur *I*.

Nous avons déjà donné ci-dessus [cf.(22) §3] l'expression de $\mathfrak{E}(S_1)$. L'expression de la variance que nous avons proposée dans le cas de la comparaison de deux valuations quelconques *X* et *Y* [Lerman 1977, 1981], nécessite l'introduction des ensembles suivants d'indéxation où des lettres différentes indiquent des indices distincts :

$$\begin{gathered} I^{[2]}, \mathrm{D} = \{[(i,j),(i,j)]\}, \ \mathrm{E} = \{[(i,j),(j,i)]\}, \\ G_1 = \{[(i,j),(i,k)]\}, G'_1 = \{[(i,j),(h,i)]\}, \\ G_2 = \{[(i,j),(h,j)]\}, G'_2 = \{[(i,j),(j,k)]\}, \\ H = \{[(i,j),(h,k)]\}. \end{gathered} \tag{3}$$

Nous désignons d'autre part par $n^{[r]} = n(n\text{-}1)...(n\text{-}r\text{+}1)$, la *r*-ème puissance factorielle de *n*. On a

$$\begin{aligned} var\,(S_1) = {} & \frac{1}{n^{[2]}}\left(\sum_{I^{([2])}} x_{ij}^2\right)\left(\sum_{I^{([2])}} y_{ij}^2\right) \\ & + \frac{1}{n^{[2]}}\left(\sum_{I^{([2])}} x_{ij}\,x_{ji}\right)\left(\sum_{I^{([2])}} y_{ij}y_{ji}\right) \\ & + \frac{1}{n^{[3]}}\left(\sum_{G_1} x_{ij}\,x_{ik}\right)\left(\sum_{G_1} y_{ij}y_{ik}\right) \\ & + \frac{1}{n^{[3]}}\left(\sum_{G'_1} x_{ij}\,x_{hi}\right)\left(\sum_{G'_1} y_{ij}y_{hi}\right) \\ & + \frac{1}{n^{[3]}}\left(\sum_{G_2} x_{ij}\,x_{hj}\right)\left(\sum_{G_2} y_{ij}y_{hj}\right) \\ & + \frac{1}{n^{[3]}}\left(\sum_{G'_2} x_{ij}\,x_{jk}\right)\left(\sum_{G'_2} y_{ij}y_{jk}\right) \\ & + \frac{1}{n^{[4]}}\left(\sum_{H} x_{ij}\,x_{hk}\right)\left(\sum_{H} y_{ij}y_{hk}\right) \\ & - \left[\frac{1}{n^{[2]}}\left(\sum_{I^{([2])}} x_{ij}\right)\left(\sum_{I^{([2])}} y_{ij}\right)\right]^2 \end{aligned} \tag{4}$$

Si on considère à présent le cas où les deux valuations *X* et *Y* sont toutes les deux, soit symétriques, soit antisymétriques [cf.(6) et (7),§ 1], l'expression ci-dessus devient

$$
\begin{aligned}
var\,(S_1) = {} & \frac{2}{n^{[2]}}\left(\sum_{I^{([2]}} x_{ij}^2\right)\left(\sum_{I^{([2]}} y_{ij}^2\right) \\
& + \frac{4}{n^{[3]}}\left(\sum_{G} x_{ij}\,x_{ik}\right)\left(\sum_{G} y_{ij}\,y_{ik}\right) \\
& + \frac{1}{n^{[4]}}\left(\sum_{H} x_{ij}\,x_{hk}\right)\left(\sum_{H} y_{ij}y_{hk}\right) \\
& - \left[\frac{1}{n^{[2]}}\left(\sum_{I^{[2]}} x_{ij}\right)\left(\sum_{I^{[2]}} y_{ij}\right)\right]
\end{aligned}
\qquad (5)
$$

où G(resp. H) est l'ensemble des tri-uplets $[i,j,k]$ (resp. quadruplées $[i,j,h,k]$) à composantes mutuellement distinctes.

Les paramètres qu'on introduit généralement pour la présentation de l'expression de var(S_1) selon N. Mantel [Mantel 1967], sont les suivants :

$$
\begin{aligned}
& A_1 = \left(\sum_{I^{([2]}} x_{ij}\right)^2 , A_2 = \sum_{i\in I}\left(\sum_{j\in I-\{i\}} x_{ij}\right)^2 , A_3 = \sum_{I^{([2]}} (x_{ij})^2 , \\
& B_1 = \left(\sum_{I^{([2]}} y_{ij}\right)^2 , B_2 = \sum_{i\in I}\left(\sum_{j\in I-\{i\}} y_{ij}\right) \text{ et } B_3 = \sum_{I^{([2]}} (y_{ij})^2
\end{aligned}
\qquad (6)
$$

Dans le cas de la comparaison de deux codages symétriques (resp. antisymétriques), on a :

$$
\begin{aligned}
var\,(S_1) = {} & \frac{2}{n^{[2]}} A_3\,B_3 + \frac{4}{n^{[3]}}(A_2 - A_3)(B_2 - B_3) \\
& + \frac{1}{n^{[4]}}(A_1 - 4A_2 + 2A_3)(B_1 - 4B_2 + 2B_3) \\
& \frac{1}{(n^{[2]})^2\,A_1B_1}
\end{aligned}
\qquad (7)
$$

La correspondance entre notre expression (5) et celle (7) de N. Mantel, peut être établie en identifiant terme à terme dans les deux expressions les facteurs de $\frac{2}{n^{[2]}}$, $\frac{4}{n^{[3]}}$, $\frac{1}{n^{[4]}}$ et $\frac{1}{(n^{[2]})^2}$.

Si l'expression de N. Mantel se prête mieux au calcul puisque les termes qu'elle comprend représentent des sommes doubles ; la nôtre, en revanche, est d'un point de vue formel, conceptuellement plus claire ; ce qui nous permettra d'étendre le calcul pour la comparaison de deux relations dont l'arité q est supérieure à 2.

4.2. *Expression "factorielle" de la variance*

Nous l'appelons ainsi parce que nous allons exprimer var(S_1) en fonction de trois paramètres de base qui sont des moments.

Commençons par introduire, relativement à $\{x_{ij} / (i,j) \in I^{[2]}$ les paramètres suivants que nous situons par rapport à ceux de N. Mantel :

$$
\begin{aligned}
& L_1 = \Sigma\,\{x(i,j)/(i,j) \in I^{[2]}\} = \sqrt{A_1} , \\
& L_2 = \Sigma\,\{(x(i,j)^2/(i,j) \in I^{[2]}\} = A_3 , \\
\text{et} \quad & U = \Sigma\left\{x\,(i,j)\,x\,(i,k) / (i,j,k) \in G\right\} = (A_2 - A_3)
\end{aligned}
\qquad (8)
$$

On introduit de même, par rapport au second codage

$$\{y_{ij} / (i,j) \in I^{[2]}\}\ ,\ M_1, M_2, et\ V, \text{ où} \qquad (9)$$
$$M_1 = \sqrt{B_1}, M_2 = B_3 \text{ et } V = (B_2 - B_3)$$

Désignons alors par :

$$\begin{aligned} &l_1 = L_1/n^{[2]}\ ,\ l_2 = L_2/n^{[2]}\ ,l_1 = u = U/n^{[3]}, \\ &m_1 = M_1/n^{[2]},\ m_2 = M_2/n^{[2]} \text{ et } v = V/n^{[2]}. \end{aligned} \qquad (10)$$

les paramètres qui sont des moments, par rapport auxquels nous allons exprimer var(S_1) en démarrant de l'expression (7) qui devient :

$$\frac{2\,L_2 M_2}{n^{[2]}} + \frac{4\,UV}{n^{[3]}} + \frac{(L_1^2 - 2L_2 - 4U)(M_1^2 - 2\,M_2 - 4\,V)}{n^{[4]}} - \frac{L_1^2 M_1^2}{(n^{[2]})^2}\ . \qquad (11)$$

L'idée générale du calcul est de commencer par substituer aux sommes L_1, L_2 et U (resp. M_1, M_2 et V) les moyennes l_1, l_2 et u (resp. m_1,m_2 et v) qui joueront le rôle de paramètres.

La taille de l'échantillon tendant vers l'infini, nous allons ensuite supposer des conditions de régularité asymptotiques qui font que ces paramètres -qui sont des moyennes- convergent presque sûrement vers des limites identifiables. D'autre part, l'expression finale du coefficient d'association est continue par rapport à ces paramètres. Il s'agira alors de décomposer var(S_1) selon des fractions rationnelles de monômes factorielles de la forme :

$$c \times \frac{(n^{[r]})^k}{(n - h)^{[s]}}\ , \qquad (12)$$

où c, r, k, h et s sont des entiers ; cette décomposition se faisant par ordre de grandeur décroissant en n ; c'est-à-dire, plus précisément, par valeurs décroissantes de kr/s. Un autre guide dans le calcul consiste à procéder à des regroupements permettant la mise en évidence de facteurs tels que $(u - l_1^2)$ [*resp.* $(v - m_1^2)$] $(l_2 - l_1^2)$ [*resp.* $((m_2 - m_1^2)$] *et* $(u - l_2)$ [*resp.* $(v - m_2)$].

Le développement tenant compte du premier principe conduit à l'expression suivante de la variance :

$$\begin{aligned} var\ (S_1) = {} & 2n^{[2]}\, l_2 m_2 + 4n^{[3]}\, uv + 4 \times \frac{(n^{[2]})^2}{n - 3}\, l_1^2\, m_1^2 \\ & + 2\,\frac{(n^{[2]})^2}{(n - 2)^{[2]}}\, l_1^2 m_1^2 + 4\ \times \frac{n^{[2]}}{(n - 2)^{[2]}}\, l_2\, m_2 \\ & + 16\,\frac{(n^{[3]})}{(n - 3)}\, uv\ - 2\ \times \frac{n^{[2]2}}{(n - 2)^{[2]}} \left(l_1^2\, m_2 + l_2 m_1^2\right) \\ & - 4\,\frac{(n^{[2]})^2}{(n - 3)}\,(l_1^2 v\ + m_1^2\, u) \\ & + 8\,\frac{(n^{[2]})}{(n - 3)}\,(l_2 v\ + m_2 u). \end{aligned} \qquad (13)$$

En obéissant aux autres principes de calcul ci-dessus exprimés et en effectuant les simplifications qui s'imposent, on obtient l'expression suivante, objet du lemme 1.

LEMME 1. En tenant compte des notations (8), (9) et (10), la variance de l'indice aléatoire S_1 [cf. (16) §3, (1) et (2) §4], peut s'écrire comme suit :

$$\begin{aligned} var\,(S_1) = \frac{2n^{[2]}}{(n-3)} \times [2n^{[2]}\,(u - l_1^2)(v - m_1^2) \\ + \frac{n^{[2]}}{(n-2)}\,(l_2 - l_1^2)(m_2 - m_1^2) \\ - 4\,(u - l_2)(v - m_2)] \end{aligned} \qquad (14)$$

4.3. Expression "absolue" de la variance et forme limite du coefficient d'association

Nous l'apelons ainsi parce que nous allons exprimer var(S_1) en fonction de trois paramètres de base qui sont des moments absolus.

Commençons par remarquer que $(l_2 - l_1^2)$ [resp. $(m_2 - m_1^2)$] représente la variance de $\{x_{ij} / (i,j) \in I^{[2]}\}$ (resp.$\{y_{ij} / (i,j)\, I^{[2]}\}$). Toutefois, la décomposition fournie par (14) n'a pas son terme dominant srictement positif. En effet, comme on va le voir $(u - l_1^2)$ peut être négatif et $(v - m_1^2)$, positif. Imaginons que x et y soient deux codages en zéros et uns correspondants aux relations d'équivalence définies par deux partitions. Nous supposons que x est associé à une partition en h classes de même cardinal $k = \frac{n}{h}$; soit $\frac{k}{n} = \frac{1}{h}$.

Dans ces conditions, on a :

$$L_1 = L_2 = h\,k\,(k-1) \text{ et } U = h\,k\,(k-1)(k-2). \text{ D'où}$$

$$l_1 = l_2 = h\,k\,(k-1) / n(n-1) \text{ et } u = h\,k\,(k-1)(k-2) / n(n-1)(n-2).$$

$$(u - l_1^2) = \frac{1}{(n-1)^2(n-2)}\left[\,(n-1)(k-1)(k-2) - (n-2)(k-1)^2\right]$$

Après développement et simplification effectués à l'intérieur du crochet, on obtient

$$\begin{aligned} (u - l_1^2 &= \frac{1}{(n-1)^2(n-2)}\left[\,n\,k + n + k^2 - k\right] \\ &= \frac{-n^2}{(n-1)^2(n-2)}\left(1 - \frac{1}{h}\right)\left(\frac{1}{h} - \frac{1}{n}\right) < 0\,. \end{aligned} \qquad (15)$$

Pour montrer une situation - d'ailleurs la plus fréquente - pour une variable partition qu'on note ici y où $(v - m_1^2)$ est positif, il suffit de prendre un exemple numérique. Imaginons $n = 90$ et une partition en deux classes de tailles respectives 60 et 30. On obtient $v = 0.3258$ et $m_1^2 = 0.3031$.

Les conditions de comportement limite s'expriment plus aisément par rapport aux moments absolus. Introduisons alors relativement à x(resp. y) la diagonale $\{x_{ii} / 1 \le i \le n\}$ (resp.$\{y_{ii} / 1 \le i \le n\}$). Rappelons [cf. (ii)§31.1.] qu'on suppose de façon exclusive l'une ou l'autre des deux situations suivantes :

$$\left.\begin{aligned} &1 : (x_{ii} = 1) \wedge (y_{ii} = 1) \text{ pour tout } i = 1, 2, ..., n \\ &2 : (x_{ii} = 0) \wedge (y_{ii} = 0) \text{ pour tout } i = 1, 2, ..., n \end{aligned}\right\} \qquad (16)$$

1 correspond au cas où les deux relations sont symétriques et 2 correspond au cas où les deux relations sont antisymétriques [cf. (6) et (7)§1]. On a

$$\sum_i x_{ii} = \sum_i x_{ii}^2$$
$$= \sum_i y_{ii} = \sum_i y_{ii}^2$$
$$= n(\text{resp.}0) \text{ dans le cas 1 (resp. 2).} \qquad (17)$$

D peut désigner la valeur commune ci-dessus (n ou 0) et $d = \frac{D}{n}$. est égal à 1 ou à 0. Introduisons :

$$\left.\begin{aligned} L_{10} = L_1 + D &= \sum_{(i,j)} \{x_{ij} / (i,j) \in I \times I\} \\ L_{20} = L_2 + D &= \sum \{x_{ij}^2 / (i,j) \in I \times I\} \\ U_0 = &\sum \{x_{ij}\, x_{ik} / (i,j,k) \in I \times I \times I\} \end{aligned}\right\} \qquad (18)$$

On a

$$U_0 = \sum_{[i,j,k]} x_{ij}\, x_{ik} + \sum_{[i,k]} x_{ii}\, x_{ik} + \sum_{[i,j]} x_{ij}\, x_{ii} + \sum_{[i,j]} x_{ij}^2 + \sum_i x_{ii}^2,$$

où un t-uplet entre crochets indique que les composantes sont mutuellement distinctes.

$$U_0 = U + 2\, d\, L_1 + L_2 + nd$$

D'où

$$\begin{aligned} U &= U_0 + 2\, d\, L_1 - L_2 - nd = U_0 - 2\, d\, (L_{10} - nd) - (L_{2\,0} - nd) - nd \\ &= U_0 - 2\, d\, L_{10} - L_{20} + 2nd^2. \end{aligned} \qquad (19)$$

Introduisons à présent les moments absolus :

$$p_1 = L_{10} / n^2,\ p_2 = L_{20} / n^2 \text{ et } q = U_0 / n^3 \qquad (20)$$

De façon correspondante nous associerons à $\{y_{ij} / (i,j) \in I \times I\}$, respectivement et avec des notations que l'on comprend :

$$r_1 = M_{10} / n^2,\ r_2 = M_{20} / n^2 \text{ et } s = V_0 / n^3 \qquad (21)$$

Avec ces paramètres, nous allons reprendre l'expression (14) de la variance.

Compte tenu de (19), on a

$$u = \frac{n^2}{(n-1)(n-2)} \left[q - \frac{2d}{n} p_1 - \frac{1}{n} p_2 + 2\, \frac{d^2}{n^2} \right], \qquad (22)$$

d'autre part,

$$l_1 = \frac{n}{(n-1)} \left(p_1 - \frac{d}{n} \right) \text{ et } l_1^2 = \frac{n^2}{(n-1)^2} \left(p_1^2 - \frac{2}{n}\, dp_1 + \frac{d^2}{n^2} \right), \qquad (23)$$

Le calcul de$(u - l_1^2)$ se simplifie pour fournir

$$(u - l_1^2) = \frac{n^2}{(n - 1(n-2))} \left(q - p_1^2 \right) - \frac{1}{(n-2)} \left(l_2\, l_1^2 \right), \qquad (24)$$

en effet, on a

$$l_2 = \frac{n}{n-1} p_2 - \frac{d}{n-1}, \qquad (25)$$

où $d = d^2 = 0$ ou 1.

On exprimera de façon semblable, v, m_1 et $(v - m_1^2)$. L'idée générale du calcul est de tout exprimer par rapport à $(q - p_1^2)$ *et* $(l_2 - l_1^2)$ [*resp.* $s - r_1^2$) *et* $(m_2 - m_1^2)$] qui, comme nous l'avons mentionné, est une variance. Ainsi, on écrira :

$$(u - l_a) = \frac{n^2}{(n-1)(n-2)}(q - p_1^2) - \frac{(n-1)}{n-2}(l_2 - l_1^2)\,; \tag{26}$$

ainsi qu'une expression correspondante pour $(v - m_2)$.

Un patient calcul conduit à l'expression exacte suivante de la variance [Lerman 87_b] :

LEMME 2.

$$\begin{aligned} var\,(S_1) &= \frac{4\,n^6}{(n-1)(n-2)^2}(q - p_1^2)(s - r_1^2) \\ &+ \frac{2n(n-1)^2}{n-3}\left[\left((l_2 - l_1^2) - \frac{2n^2}{(n-1)(n-2)}(q - p_1^2)\right)\right] \\ &\times\left[\left((m_2 - m_1^2) - \frac{2n^2}{(n-1)(n-2)}(s - r_1^2)\right)\right] \end{aligned} \tag{27}$$

THÉORÈME 1. Dans des conditions asymptotiques où p_1, p_2 et q (resp. r_1, r_2 et s) tendent vers des limites finies, pour n tendant vers l'infini, la forme limite, pour n "grand", de var(S_1) est :

$$4n^3\left\{(q - p_1^2)(s - r_1^2) + \frac{1}{2n}[(p_2 - p_1^2) - 2(q - p_1^2)] \times [(r_2 - r_2^1) - 2(s - r_1^2)]\right\} \tag{28}$$

PROPRIÉTÉ. La décomposition ci-dessus de var(S_1) est en éléments positifs.

La démonstration résultera d'une interprétation très précise de la structure de $(q - p_1^2)$ et de $[(p_2 - p_1^2) - 2(q - p_1^2)]$ que nous aurions pu écrire sous la forme

$$[(p_2 - q) - (q - p_1^2)] \tag{29}$$

Il en sera bien sûr, respectivement de même pour l'interprétation de $(s - r_1^2)$ *et de* $[(r_2 - s) - (s - r_1^2)]$.

Si on considère sur I, la variable dont la valeur $x_{i.}$ sur i, représente la moyenne de la valuation x sur l'ensemble des arcs d'origine i :

$$x_{i.} = \frac{1}{n}\sum_{1 \le j \le n} x_{ij}\,, \tag{30}$$

$(q - p_1^2)$ représente exactement la variance d'une telle variable. Ainsi,

$$q - p_1^2 = \frac{1}{n}\sum_{1 \le i \le n}(x_{i.} - x_{..})^2\,, \tag{31}$$

où

$$x_{..} = \frac{1}{n}\sum_{1 \le i \le n} x_{i.}$$

D'autre part, $(p_2 - q)$ peut s'écrire sous la forme :

$$p_2 - q = \frac{1}{n}\sum_i \left[\frac{1}{n}\sum_j x_{ij}^2 - \left(\frac{1}{n}\sum_j x_{ij}\right)^2\right] = \frac{1}{n}\sum_i \left[\frac{1}{n}\sum_j (x_{ij} - x_{i.})^2\right], \tag{32}$$

qui représente la moyenne des variances de la suite de n distributions, où la i-ème définie par $\{x_{ij} / 1 \leq j \leq n\}$, représente la suite des valuations des arcs issus de i, $1 \leq i \leq n$.

$(p_2 - q)$ a le sens d'une variance *intra* et $(q - p_1^2)$ celui d'une variance *inter*, par rapport à la décomposition de la variance générale :

$$p_2 - p_1^2 = (p_2 - q) + (q - p_1)^2. \tag{33}$$

Nous démontrerons ci-dessous, dans deux cas qualitatifs particulièrement importants que

$$p_2 - q > q - p_1^2. \tag{34}$$

Mais nous n'avons pas encore pu *directement* démontrer l'inégalité (34) en toute généralité, en partant de l'hypothèse que la valuation X est symétrique (resp. antisymétrique). Ce résultat sera ici une conséquence de la positivité de $var(S_1)$.

Imaginons en effet que la valuation Y soit telle que la moyenne $y_{i.}$ soit une constante par rapport i :

$$y_{i.} = \frac{1}{n}\sum_{1 \leq j \leq n} y_{ij} = \text{constante} \tag{35}$$

pour tout $i = 1,2,...,n$.

Pour fixer les idées dans le cas symétrique, on peut considérer l'exemple où le codage y est celui d'une relation de partition en classes de même cardinal. Pour illustrer le cas antisymétrique, on peut imaginer une répartition uniforme de n points sur un cercle orienté (par exemple dans le sens des aiguilles d'une montre), où on pose $y_{ij} = \varepsilon\, d_{ij}$, où d_{ij} est la longueur de l'arc le plus court séparant i de j et où $\varepsilon = +1$ (resp. -1) si cet arc orienté va de i vers j (resp. de j vers i).

Dans la situation définie par (35) le terme dominant de $var(S_1)$ est nul en raison de la nullité de $(s - r_1^2)$. Le reste se réduit à

$$\frac{1}{2n}\left[(p_2 - q) - (q - p_1^2)\right](r_2 - r_1^2). \tag{36}$$

Or $(r_2 - r_1^2)$ qui est une variance, est *strictement* positif ; sinon, on se retrouve dans un cas dégénéré où la valuation y est constante et que nous écartons. Il s'agit en effet d'une relation qui ne discrimine rien. Donc, on a nécessairement

$$p_2 - q > q - p_1^2 \tag{37}$$

Nous allons à présent déterminer la forme limite du coefficient $Q(X,Y)$ [cf. (19)§3.2]. Le numérateur de ce coefficient peut se mettre sous la forme :

$$n^{[2]}\,(\gamma - l_1\, m_1)\,, \tag{38}$$

où

$$\gamma = \frac{1}{n^{[2]}} \times \Sigma \left\{ x_{ij}\, y_{ij} / (i,j) \in I^{[2]} \right\}$$

La taille n de l'échantillon tendant vers l'infini, nous supposons des conditions de régularité asymptotique pour lesquelles γ, l_1 et m_1 convergent presque sûrement vers des limites identifiables. La convergence de γ est équivalente à celle de

$$w = \frac{1}{n^{[2]}} \Sigma \left\{ x_{ij}\, y_{ij} / (i,j) \in I \times I \right\} \qquad (39)$$

où on aura chargé la diagonale de $I \times I$ conformément à (16) ci-dessus. Ainsi, l'expression limite de (38) comporte comme terme dominant

$$n^2(w - p_1 r_1) \qquad (40)$$

THÉORÈME 2. La forme limite du coefficient $Q(X,Y)$ pour w ainsi que p_1,p_2 et q (resp. r_1,r_2 et s) tendant presque sûrement vers des limites finies, lorsque la taille n de l'échantillon aléatoire O tend vers l'infini, est :

$$\frac{\sqrt{n}}{2} \times \frac{(w - p_1 r_1)}{\sqrt{(q - p_1^2)(s - r_1^2) + \frac{1}{2n}\left[(p_2 - p_1^2) - 2(q - p_1^2)\right]\left[(r_2 - r_1^2) - 2(s - r_1^2)\right]}}. \qquad (41)$$

Si $(q - p_1^2)(s - r_1^2)$ est différent de zéro et si n est assez grand pour rendre négligeable le deuxième terme sous le signe dénominateur, on a :

$$Q(X,Y) \cong \frac{\sqrt{n}}{2} \times \frac{(w - p_1 r_1)}{\sqrt{(q - p_1^2)(s - r_1^2)}}. \qquad (42)$$

On se trouve alors en mesure de répondre à la question posée à la fin du paragraphe 3.2. Comme pour celui de K.Pearson, le coefficient Q(X,Y) entre deux variables relationnelles est également -dans les conditions du théorème 2, pour $(q\text{-}p^2{}_1)(s\text{-}r^2{}_1) \neq 0$, au facteur $\sqrt{n}$ près- un rapport pur dont la limite est indépendante de n.

Comme nous pourrons le voir dans le cas de variables partition ou préordre total $(q\text{-}p^2{}_1)$ [resp.$(s\text{-}r^2{}_1)$] est nécessairement positif, compte tenu de l'expression (28) de la variance.

5. APPLICATIONS AUX CAS QUALITATIFS NOMINAL ET ORDINAL

5.1. Le cas nominal

5.1.1. Introduction

Par rapport au schéma de la figure 1 du paragraphe 1, les deux structures α et β, respectivement définies sur l'ensemble O des objets par les deux variables qualitatives nominales à comparer, sont deux partitions que nous noterons π et χ ; plus précisément, on pourra écrire :

$$\pi = \{E_i / 1 \leq i \leq h\}$$

$$\left(\text{resp. } \chi = \{F_j / 1 \leq j \leq k\}\right) \qquad (1)$$

où E_i (resp. F_j) est la i-ème (resp. j-ème) classe de la partition π (resp. χ), qu'on supposera -

sans restreindre la généralité - qu'elle est en classes étiquetées, $1 \leq i \leq k$ (resp. $1 \leq j \leq k$).

$$t\ (\pi) = (m_i / 1 \leq i \leq h)$$
$$\left[\text{resp. } t(\chi) = (n_j / 1 \leq j \leq k)\right], \qquad (2)$$

où m_i = card (E_i) [resp. n_j = card (F_j)].

Dans [Lerman 1973,1981], nous avons raisonné dans un contexte purement ensembliste et dans le cadre de l'hypothèse d'absence de liaison H_1 (cf. §3.1.). Nous introduisons l'ensemble de représentation $O^{\{2\}}$ et représentons une partition par l'ensemble des paires qu'elle réunit [cf. (2) §1]. Ainsi, avec des notations que l'on comprend,

$$R(\pi) = \sum \left\{E_i^{\{2\}} / 1 \leq i \leq h\right\}$$
$$\left[\text{resp. } R(\chi) = \sum \{F_j^{\{2\}} / 1 \leq j \leq k\}\right] \qquad (3)$$

Ω_π (resp. Ω_χ) [cf. schéma de la figure 1 du paragraphe 1] peut directement être défini comme étant l'ensemble des parties de $O^{\{2\}}$ dont chacune correspond à la représentation d'une partition de type $t(\pi)$ [resp. $t(\chi)$]. On a, en désignant par $\pi \wedge \chi$, la partition résultant du croisement des deux partitions π et χ :

$$\begin{aligned} s(\pi,\chi) &= \text{card}\,[R(\pi) \cap R(\chi)] = \text{card}\,[R(\pi \wedge \chi)] \\ &= \text{card}\left[\sum \left\{(E_i \cap F_j)^{\{2\}} / 1 \leq i \leq h, 1 \leq j \leq k\right\}\right] \\ &= \sum \left\{n_{ij}\ (n_{ij} - 1)/2 \,/\, 1 \leq i \leq h, 1 \leq j \leq k\right\}, \end{aligned} \qquad (4)$$

où nous avons noté $\pi \wedge \chi$ le croisement des deux partitions π et χ et n_{ij} le cardinal de $E_i \cap F_j$, $1 \leq i \leq h$, $1 \leq j \leq k$.

Conformément aux notations du paragaphe 1, nous avions obtenu les expressions suivantes :

$$E\left[s(\pi^*,\chi^*)\right] = \lambda\mu \text{ et } \text{var}\left[s(\pi^*,\chi^*)\right] = \lambda\mu + \rho\sigma + \theta\zeta - \lambda^2\mu^2.. \qquad (5)$$

où
$$\lambda = \sum \left\{m_i(m_i - 1)/\sqrt{2n(n-1)} / 1 \leq i \leq h\right\},$$
$$\rho = \sum \left\{m_i(m_i - 1)(m_i - 2)/\sqrt{n(n-1)(n-2)} \,/\, 1 \leq i \leq h\right\},$$
$$\theta = \left\{\left[\sum_i m_i(m_i - 1)\right]^2 - 2\sum_i m_i(m_i - 1)(2m_i - 3)\right\} / 2\sqrt{n(n-1)(n-2)(n-3)}$$

et où les expressions de μ, σ et ζ ont respectivement la même forme que λ, ρ et θ ; les m_i de $t(\tau)$, étant remplacés par les n_j de $t(\chi)$, $1 \leq j \leq k$.

Rappelons que la distribution de $s(\pi^*,\chi^*)$, où π^* et χ^* sont deux partitions aléatoires indépendantes, est la même que celle de $s(\pi^*,\chi)$[resp. $s(\pi,\chi^*)$]. Dans [Daudé 1992] on trouvera l'extension de l'étude de l'indice aléatoire $s(\pi^*,\chi^*)$ dans le cas des hypothèses d'absence de liaison H_2 et H_3.

L'indice qui nous intéresse est, comme toujours, celui centré et réduit :

$$Q_1(\pi,\chi) = \frac{s(\pi,\chi) - E\left[s(\pi^*,\chi^*)\right]}{\sqrt{\mathrm{var}\left[s(\pi^*,\chi^*)\right]}} \qquad (6)$$

5.1.2. Forme limite de $Q_1(\pi,\chi)$.

Commençons par préciser que les indices courants $i,j,k,...$, indiqueront ici des éléments de l'ensemble O des objets, alors qu'au paragraphe 5.1.1. précédent, ils correspondaient à des étiquettes de classes.

En codant la partition π(resp. χ) par une valuation $X = \{x_{ij} / (i,j) \in I \times I\}$(resp.$Y = \{y_{ij} / (i,j) \in I \times I\}$), telle que x_{ij} (resp. y_{ij}) est égal à 1 ou à 0, selon que i et j sont ou non reliés par la relation d'équivalence définie par π(resp.χ) ; on obtient exactement le même coefficient (6) ci-dessus, au moyen de l'indice $Q_1(X,Y)$, élaboré sous l'hypothèse d'absence de liaison H_1 [cf.§3.2.]. La raison en est l'invariance de la valuation sur la diagonale de $I \times I$; ici on a : $x_{ii} = 1$ (resp.$y_{ii} = 1$) pour tout $i = 1,2,...,n$. Dans ces conditions, pour déterminer la forme limite du coefficient d'association $Q_1(\pi,\chi)$ [cf. (6) ci-dessus], il suffit d'appliquer les résultats du paragraphe 5 précédent en leur donnant un accent particulier ; c'est-à-dire, de se rendre compte de ce que deviennent chacun des paramètres p_1, p_2, q (resp. r_1, r_2, s) et w. On notera ici par c (resp. d) l'indice courant, étiquette d'une classe de la partition π (resp.χ) de type $t(\pi) = (m_1, m_2, ..., m_c, ..., m_g)$ [resp. $(n_1, n_2, ..., n_d, ..., n_h)$]. Désignons par $\pi_c = m_c / n$ [resp. $\chi_d = n_d / n$] et $\nu_{cd} = n_{cd}$, où n_{cd} est le cardinal de la classe (c,d) de $\pi \wedge \chi$. On a alors

PROPRIÉTÉ 1.

$$\begin{aligned} p_1 = p_2 &= \sum_{1 \le c \le g} \pi_c^2 \,,\; q = \sum_{1 \le c \le g} \pi_c^3 \\ (resp.\; r_1 = r_2 &= \sum_{1 \le d \le h} \chi_d^2 \,,\; s = \sum_{1 \le d \le h} \chi_d^3 \\ \text{et } w &= \Sigma \left\{ \nu_{cd}^2 / \, 1 \le c \le g \,,\, 1 \le d \le h \right\} \end{aligned} \qquad (7)$$

En effet, x_{ij} étant une variable booléenne

$$\sum_{(i,j)} x_{ij} = \sum_{(i,j)} x_{ij}^2 = \sum_{(1 \le c \le g)} m_c^2 \,, \qquad (8)$$

puisque le membre de gauche représente le nombre de couples d'objets réunis par la partition π $p = p_1 = p_2$ représente la proportion de couples d'objets réunis par π.

D'autre part $x_{ij}\, x_{ik}$ n'est égal à 1 que si i,j et k se retrouvent dans une même classe. Le nombre de triplets de la classe E_c de cardinal m_c constitue la contribution de cette classe à $\Sigma\{x_{ij}\, x_{ik} / (i,j,k)\}$, d'où la relation :

$$\sum_{(i,j)} \{x_{ij}\, x_{ik} \,/\, (i,j,k)\} = \sum_{(1 \le c \le g)} m_c^3 \,, \qquad (9)$$

il en résulte l'expression de q. Quant à w, on a

$$\sum_{(i,j)} \{x_{ij}\, y_{ik} / (i,j) \in I \times I\} = \Sigma\{n_{cd}^2 / 1 \leq c \leq g \;\; 1 \leq d \leq h\}, \qquad (10)$$

puisque le premier membre représente le nombre de couples d'objets réunis par la partition croisée $\pi \wedge \chi$. D'où w.

Dans ces conditions, on peut énoncer une conséquence directe du théorème 2 du paragraphe précédent.

PROPRIÉTÉ 2. La taille n de l'échantillon aléatoire O tendant vers l'infini de telle sorte que les proportions empiriques $\{v_{cd}/1 \leq c \leq g,\ 1 \leq d \leq h\}$ convergent presque sûrement vers leurs limites théoriques $\{\tau_{cd} / 1 \leq c \leq g,\ 1 \leq d \leq h\}$, la forme limite de $Q_1(\pi,\chi)$ est :

$$\frac{\sqrt{n}}{2} \times \frac{(w - pr)}{\sqrt{(q - p^2)(s - r^2) + \frac{1}{2n}[(p - p^2) - 2(q - p^2)][(r - r^2) - 2(s - r^2)]}} \qquad (11)$$

où w, p (resp. r) et q(resp. s) tendent, presque sûrement, vers

$$\Sigma\{\tau_{cd}^2 / 1 \leq c \leq g\ ,\ 1 \leq d \leq h\},\quad \Sigma\{\tau_{c.}^2 / 1 \leq c \leq g\}$$
$$\left(\text{resp}\quad \Sigma\{\tau_{.d}^2 / 1 \leq d \leq h\} \text{ et } \Sigma\{\tau_{c.}^3 / 1 \leq c \leq g\}\right)$$
$$\left(\text{resp}\quad \Sigma\{\tau_{.d}^3 / 1 \leq d \leq h\} \text{ et } \tau_{c.}\ (\text{resp.}\tau_{.d})\ ,\right)$$

limite de π_c(resp. χ_d) est la somme pour d (resp. c) des τ_{cd}.

Nous allons directement établir la

PROPRIÉTÉ 3. $(q - p^2)$ est nul si $\pi_c = 1/g$ pour tout $c = 1,2,...,g$, autrement $(q - p^2)$ est strictement positif.

$$\text{Si} \qquad \pi_c = 1/g \text{ pour tout } c = 1,2,...,g,$$
$$p = \sum_{1 \leq c \leq g} \frac{1}{g^2} = \frac{1}{g} \text{ et } p^2 = \frac{1}{g^2},$$
$$q = \sum_{1 \leq c \leq g} \frac{1}{g^3} = \frac{1}{g^2}.$$

Prouvons à présent que

$$\sum_{1 \leq c \leq g} \pi_c^3 \geq \left(\sum_{1 \leq c \leq g} \pi_c^2\right)^2 \qquad (12)$$

Cette inégalité est équivalente à la suivante

$$\left(\sum_{1 \leq c \leq g} \pi_c^3\right)\left(\sum_{1 \leq c \leq g} \pi_c\right) \geq \sum_{c} \pi_c^4 + \sum_{(c,c')} \pi_c^2\, \pi_{c'}^2\ , \qquad (13)$$

où la dernière somme a lieu pour tous les couples (c,c') à composantes distinctes. Le développement du premier membre laisse à prouver

$$\sum_{[c,c']} \pi_c^3\, \pi_{c'} \geq \sum_{[c,c']} \pi_c^2\, \pi_{c'}^2, \qquad (14)$$

L'inégalité sera *a fortiori* établie si chaque couple [c,c'] contribue à l'inégalité ; c'est-à-dire, si

$$\pi_c^3\,\pi_{c'} + \pi_{c'}^3\,,\,\pi_c \geq 2\pi_c^2\,\pi_{c'}^2\,, \tag{15}$$

ce qu'on voit aisément en divisant les deux membres par $\pi_c\,\pi_{c'}$. D'autre part, il suffit que π_c soit différent de $\pi_{c'}$ pour un [c,c'], pour que l'inégalité (15) et donc celle (12) soit stricte.

PROPRIÉTÉ 4. $A = [(p - p^2) - 2(q - p^2)]$. Compte tenu de la décomposition sous-jacente à la preuve de la propriété 2

$$(q - p^2) = \sum_{\{c,c'\}} \pi_c\,\pi_{c'}\,(\pi_c - \pi_{c'})^2. \tag{16}$$

où la somme a lieu pour les $\binom{g}{2}$ paires $\{c,c'\}$. On suppose -sans restreindre la généralité- que c est tel que $\pi_c \geq \pi_{c'}$.

$$(p - q) = \sum_{c} \pi_c^2\,(1 - \pi_c) = \sum_{[c,c']} \pi_c^2\,\pi_c = \sum_{\{c,c'\}} \pi_c\,\pi_{c'}\,(\pi_c + \pi_{c'}). \tag{17}$$

Or, pour tout $\{c,c'\}$,

$$\pi_c + \pi_{c'} = (\pi_c - \pi_{c'}) + 2\,\pi_{c'} > (\pi_c - \pi_{c'})^2 \tag{18}$$

et

$$(p - q) > (q - p^2). \tag{19}$$

On remarquera qu'à l'extrémum, où $\pi_c = \frac{1}{g}$ pour tout c, la valeur de A est $\frac{1}{g}\,(1 - \frac{1}{g})$. Ainsi la décomposition fournie de la variance [sous le signe $\sqrt{\ }$ de (11)] est en éléments positifs.

5.1.3. Remarques générales sur le coefficient $Q_1(\pi,\chi)$.

Toutes choses égales par ailleurs, le coefficient $Q_1(\pi,\chi)$ est d'autant plus grand que la partition π(resp. χ) tend à être en classes de même taille ; c'est-à-dire, à être plus discriminante ou plus informative en termes de théorie de l'Information. On a même l'impression d'une rupture dans le comportement du coefficient dès lors que l'une des partitions est en classes de même effectif. Toutefois, deux remarques s'imposent qui atténuent en partie cette impression. La première est que l'ordre de grandeur du numérateur -lié à $\Sigma\left\{n_{cd}^2 / 1 \leq c \leq g\,,\,1 \leq d \leq h\right\}$ [cf. (4) ci-dessus]- décroît sensiblement dès lors que la partition π(resp. χ) tend à être en classes de même cardinal. La seconde remarque est que les partitions en classes de même taille ne se rencontrent pas naturellement, elles peuvent correspondre à un échantillonnage sous contraintes et sont, par conséquent, de pures constructions.

Ainsi, le coefficient $Q_1(\pi, \chi) / \sqrt{n}$ se trouve marqué par la valeur informative de chacune des deux partitions π et χ. Il s'agit d'une tendance qui peut être souhaitée. Si maintenant, on désire un coefficient masquant cet effet et ne mettant en évidence que la similarité des formes, on pourra prendre

$$R = (\pi\,,\,\chi) = \frac{Q_1(\pi\,,\,\chi)}{\sqrt{Q_1(\pi,\pi)\,,Q_1(\chi,\chi)}} \tag{20}$$

qui se réfère à (21) du paragraphe 3.2..

On pourra remarquer que, dans des conditions assez générales où la partition π (resp. χ) ne tend pas -pour n assez grand- vers une partition en classes de même taille, $R(\pi, \chi)$ tend vers le coefficient de K. Pearson [cf. (23) de 2.2. (i)] entre les deux attributs booléens définis au niveau de $O^{(2)}$ et respectivement représentés par $R(\pi)$ et $R(\chi)$ [cf. (3) ci-dessus].

5.2. LE CAS ORDINAL

5.2.1. Introduction

Nous nous proposons ici d'étudier le coefficient d'association obtenu par application de l'expression générale (41) du théorème 2 du paragraphe 4, dans le cas de la comparaison de deux variables qualitatives ordinales que nous allons représenter comme deux relations antisymétriques particulières sur $I \times I$, où $I = \{1,2,...,i,...,n\}$ code l'ensemble O des objets [cf. (7) §1]. Plus précisément, on peut noter ω et $\overline{\omega}$ les deux préordres totaux sur O respectivement induits par les deux variables qualitatives ordinales et poser, relativement à la valuation X (resp. Y) représentant ω (resp. $\overline{\omega}$) :

x_{ij} (resp.y_{ij}) = 1,0 ou -1, selon que i précède strictement j, est dans la même classe que j ou suit strictement j, pour le préordre ω (resp. $\overline{\omega}$).

Nous noterons ($E_c/1 \leq c \leq h$) [resp.($F_d/1 \leq d \leq k$)] la suite ordonnée des classes de ω (resp. $\overline{\omega}$). D'autre part, I_c (resp. J_d) est la partie de I codant E_c (resp. F_d), $1 \leq c \leq h$ (resp.($1 \leq d \leq k$). Enfin, $(m_1,m_2,...,m_c,...,m_h)$ [resp.$(n_1,n_2,...,n_d,...,n_k)$ indiquera la composition du préordre ω (resp. $\overline{\omega}$). Par ailleurs, on considérera les proportions suivantes : $\omega_c = m_c / n$, $\overline{\omega}_d = n_d/n$ et $\upsilon_{cd} = n_{cd}/n$, où n_{cd} est le cardinal de la classe $E_c \cap F_d$, $1 \leq c \leq h$, $1 \leq d \leq k$.

5.2.2. Forme limite de $Q_1(\omega,\overline{\omega})$.

Pour pouvoir expliciter le coefficient d'association [cf. (41) §4], il y a lieu de pouvoir reconnaître p_1,p_2,q (resp. r_1,r_2,s) et w.

PROPRIÉTÉ 1.

$$p_1 = 0, p_2 = 2 \sum_{c<c'} \omega_c\omega_{c'}, \quad q = [2 \sum_{c<c'<c''} \omega_c\omega_{c'}\omega_{c''} + \sum_{c<c'} \omega_c\omega_{c'}(\omega_c + \omega_{c'})]$$

$$(\text{resp. } r_1 = 0, r_2 = 2 \sum_{d<d} \overline{\omega}_d\overline{\omega}_{d'}, s = [2 \sum_{d<d<d'} \overline{\omega}_d\overline{\omega}_{d'}\overline{\omega}_{d''} + \sum_{d<d} \overline{\omega}_d\overline{\omega}_{d'}(\overline{\omega}_d + \overline{\omega}_{d'})])$$

et
$$w = 2 \sum_{c<c'} \sum_{d<d} (\upsilon_{cd}\upsilon_{c'd'} - \upsilon_{cd'}\upsilon_{c'd}). \qquad (21)$$

Considérons $\Sigma \{x_{ij}/(i,j) \in I \times I\}$ et la décomposition suivante de $I \times I$ conformément à ω :

$$I \times I = \sum_{c<c'} I_c \times I_{c'} + \sum_{c'>c} I_{c'} \times I_c + \sum_{c} I_c \times I_c \qquad (22)$$

(somme ensembliste)

Compte tenu du codage antisymétrique, $I_c \times I_c$ contribue pour zéro à la somme. D'autre part, pour c et c' fixés tels que $c < c'$, $(I_c \times I_{c'} + I_{c'} \times I_c)$ contribue pour zéro à la somme, car lorsque (i,j) décrit $I_c \times I_{c'}$, (j,i) décrit $I_{c'} \times I_c$ et on a constamment $(x_{ij}+x_{ji}) = 0$; d'où la valeur de p_1.

Le calcul de p_2 suppose celui de la somme

$$\Sigma \{x_{ij}^2 / (i,j) \in I \times I\} \tag{23}$$

et on considère la même décomposition (22). Cette fois ci, pour c et c' fixés tels que $c < c'$, $I_c \times I_{c'}$ et $I_{c'} \times I_c$ contribuent chacun pour $m_c \times m_{c'}$; d'où le résultat annoncé pour p_2.

Considérons à présent une expresion de la forme

$$\Sigma \{x_{ij} \times x_{ik} / (i,j;k) \in I \times I \times I\} \tag{24}$$

et référons nous à la décomposition de $I \times I \times I$ conforme à celle $\{I_c / 1 \leq c \leq h\}$ de I. Un élément de la décomposition est de la forme générale $I_c \times I_{c'} \times I_{c''}$.

Considérons le triplet (c,c',c''). Si ces trois composantes sont égales, l'élément de la décomposition a la forme I_c^3. Si maintenant deux des composantes sont égales, on a pour c et c' fixés $(c < c')$ les configurations suivantes :

$$I_c^2 \times I_{c'}, I_c \times I_{c'} \times I_c, I_{c'} \times I_c^2$$

$$I_{c'}^2 \times I_c, I_{c'} \times I_c \times I_{c'}, I_c \times I_{c'}^2 .$$

Les seuls ensembles qui contribuent à (24) sont $I_{c'} \times I_c^2$ et $I_c \times I_{c'}^2$. Leur contribution est

$$\sum_{c<c'} m_c m_c (m_c-1) + \sum_{c<c'} m_c m_{c'} (m_{c'}-1) \tag{25}$$

qui contribue à l'expression de q par

$$\sum_{c<c'} \omega_c \omega_{c'} (\omega_c + \omega_{c'}) . \tag{26}$$

Considérons à présent le cas où les trois composantes c,c' et c'' sont distinctes. Nous avons à envisager les produits cartésiens à gauche ci-dessous et leurs contributions respectives, à droite ci-dessous. On suppose $c < c' < c''$.

$$I_c \times I_{c'} \times I_{c''} \rightarrow m_c m_c m_{c''}$$

$$I_c \times I_{c''} \times I_{c'} \rightarrow m_c m_c m_{c''}$$

$$I_{c'} \times I_c \times I_{c''} \rightarrow -m_c m_c m_{c''}$$

$$I_{c'} \times I_{c''} \times I_c \rightarrow -m_c m_c m_{c''}$$

$$I_{c''} \times I_c \times I_{c'} \rightarrow m_c m_c m_{c''}$$

$$I_{c''} \times I_{c'} \times I_c \rightarrow m_c m_{c'} m_{c''}$$

La contribution globale est donc

$$2 \sum_{c<c'<c''} m_c m_{c'} m_{c''} \tag{27}$$

laquelle donne au niveau de q :

$$2 \sum_{c<c'<c''} \omega_c \omega_{c'} \omega_{c''} . \tag{28}$$

Il reste maintenant à déterminer w.

Considérons $\Sigma \{x_{ij} y_{ij} / (i,j) \in I \times I\}$ et la décomposition de même type que (22), mais relative à ω :

$$I \times I = \sum_{d<d''} J_d \times J_{d'} + \sum_{d>d''} J_d \times J_{d'} + \sum_d J_d \times J_d . \tag{29}$$

Si $(i,j) \in I_c \times I_c$ pour un c, ou $(i,j) \in J_d \times J_d$ pour un d, $x_{ij} y_{ij} = 0$. Étant donné maintenant un couple de paires d'indices $(\{c,c'\},\{d,d'\})$ où $c < c'$ et $d < d'$, considérons les contributions suivantes notées à droite des ensembles suivants notés à gauche :

$$(I_c \times I_{c'}) \cap (J_d \times J_{d'}) \rightarrow n_{cd} n_{c'd'}$$

$$(I_c \times I_{c'}) \cap (J_{d'} \times J_d) \rightarrow -n_{cd'} n_{c'd}$$

$$(I_{c'} \times I_c) \cap (J_d \times J_{d'}) \rightarrow -n_{c'd} n_{cd'}$$

$$(I_{c'} \times I_c) \cap (J_{d'} \times J_d) \rightarrow n_{c'd} n_{cd} .$$

La contribution globale est

$$2n_{cd} n_{c'd'} - 2n_{cd'} n_{c'd} \tag{30}$$

Il en résulte l'expression de w donnée dans l'énoncé de la propriété.

L'expression (28) du paragraphe 4.3. de la variance devient -au facteur $4n^3$ près -

$$qs + \frac{1}{2n} (p_2 - 2q) (r_2 - 2s) \tag{31}$$

où q (resp. s) est essentiellement positif.

PROPRIÉTÉ 2. Les vecteurs $(\omega_c / 1 \leq c \leq h)$ et $(\overline{\omega}_d / 1 \leq d \leq k)$ tendant vers des vecteurs limites finis, la forme limite du coefficient d'association $Q_1(\omega,\omega)$ est :

$$\frac{\sqrt{n}}{2} \times \frac{w}{\sqrt{qs + \frac{1}{2n} (p_2 - 2q) (r_2 - 2s)}} \tag{32}$$

où les paramètres de ce coefficient ont été précisés dans la propriété 1 ci-dessus.

PROPRIÉTÉ 3. $B = (p_2 - 2q)$ est positif.

On peut écrire p_2 sous la forme

$$2 \sum_{c<c'} \omega_c \omega_{c'} \left(\sum_{c''} \omega_{c''} \right) = 2 \sum_{c<c'} \omega_c \omega_{c'} (\omega_c + \omega_{c'}) + 2 \sum_{c<c'} \omega_c \omega_{c'} (\Sigma \{\omega_{c''} / c'' \notin \{c,c'\}\}) \quad (33)$$

$$= 2 \sum_{c<c'} \omega_c \omega_{c'} (\omega_c + \omega_{c'}) + 2 \sum_{c''<c<c'} \omega_{c''} \omega_c \omega_{c'} + 2 \sum_{c<c''<c'} \omega_c \omega_{c''} \omega_{c'} + 2 \sum_{c<c'<c''} \omega_c \omega_{c'} \omega_{c''} . \quad (34)$$

Chaque terme $\omega_{c_1} \omega_{c_2} \omega_{c_3}$ — pour $c_1 < c_2 < c_3$ — se retrouve trois fois, une dans chacune des trois dernières sommes. Il en résulte que

$$B = p_2 - 2q = 2 \sum_{c<c'<c''} \omega_c \omega_{c'} \omega_{c''} > 0 \quad (35)$$

Ainsi la décomposition (31) est en éléments positifs.

Ici encore on peut confronter le comportement de $Q_1 (\omega,\bar{\omega})$ à celui de

$$R(\omega,\bar{\omega}) = \frac{Q_1 (\omega,\bar{\omega})}{\sqrt{Q_1 (\omega,\omega)\, Q_1 (\bar{\omega},\bar{\omega})}} \quad (36)$$

5.2.3. *Interprétation ensembliste et statistique du coefficient $Q_1(\omega,\bar{\omega})$ Comparaison avec un précédent coefficient.*

Nous allons commencer par donner une vision ensembliste de l'indice brut

$$s\ (X,Y) = \sum \left\{ x_{ij}\, y_{ij} / (i,j) \in I^{[2]} \right\}, \quad (37)$$

ayant abouti à la construction du coefficient $Q_1(\omega,\bar{\omega})$. A cette fin et avec des notations que l'on comprend, introduisons relativement à ω(resp. $\bar{\omega}$) les ensembles suivants :

$$\begin{aligned} A_1 &= \sum_{c<c'} \mathrm{E}_c \times \mathrm{E}_{c'} \quad (\text{resp.} B_1 = \sum_{d<d'} \mathrm{F}_d \times \mathrm{F}_{d'}) , \\ A_2 &= \sum_{c} \mathrm{E}_c^{[2]} \quad (\text{resp.} B_2 = \sum_{d} \mathrm{F}_d^{[2]}) \qquad \text{et} \\ A_3 &= \sum_{c>c'} \mathrm{E}_c \times \mathrm{E}_{c'} \quad (\text{resp.} B_3 = \sum_{d>d'} \mathrm{F}_d \times \mathrm{F}_{d'}) . \end{aligned} \quad (38)$$

Posons alors les neuf paramètres suivants :

$$\begin{aligned} &s = \mathrm{card}(A_1 \cap B_1) , \ q = \mathrm{card}(A_1 \cap B_2) , \ u = \mathrm{card}(A_1 \cap B_3), \\ &p = \mathrm{card}(A_2 \cap B_1) , \ r = \mathrm{card}(A_2 \cap B_2) , \ p' = \mathrm{card}(A_2 \cap B_3), \\ &v = \mathrm{card}(A_3 \cap B_1) , \ q' = \mathrm{card}(A_3 \cap . B_2) , \ t = \mathrm{card}(A_3 \cap B_3), \end{aligned} \quad (39)$$

où on a :

$$s = t , \ q = q' , \ u = v \ \text{ et } \ p = p' \quad (40)$$

L'indice $s(X,Y)$ [cf. (37)] peut se mettre sous la forme :

$$s(X,Y) = 2(s - u) \quad (41)$$

Il correspond exactement au numérateur du coefficient que propose M.G. Kendall [Kendall 1970] ou de celui γ de L.O. Goodman et W.H. Kruskal [Goodman et Kruskal 1954]. Nous avons longtemps pu croire à partir d'un exemple construit dans une situation de parfaite

indépendance entre les deux préordres totaux ω et $\bar{\omega}$, que l'indice $(s - u)$ est biaisé par rapport à l'hypothèse d'absence de liaison H_1 définie au paragraphe 2 ci-dessus [Lerman 1973,1981b]. Nous venons de nous rendre compte que la conception de ce calcul est erronée (*errare humanum est*) et qu'en fait, on a très précisément :

$$\mathcal{E}\,(s^*) = \mathcal{E}\,(u^*) = \frac{1}{n[2]}\left(\sum_{c<c'} m_c\, m_{c'}\right)\left(\sum_{d<d'} n_d\, n_{d'}\right) \tag{42}$$

Le coefficient que nous proposons s'écrit dans ces conditions :

$$Q_1(\omega,\bar{\omega}) = \frac{s - u}{\sqrt{\text{var}\,(s^* - u^*)}} \tag{43}$$

Dans notre précédente approche, le point d'ancrage a consisté dans la comparaison d'un couple de variables "rang" définissant un couple $(\omega,\bar{\omega})$ d'ordres totaux et stricts sur O. Dans ce cas $R(\omega)$ [resp. $R(\bar{\omega})$] est le graphe dans $O \times O$ de la relation d'ordre stricte définie par ω (resp. $\bar{\omega}$). Dans ce dernier cas, certains des ensembles impliqués dans les expressions (39) deviennent vides ; et, on a alors :

$$q = p = r = p' = q' = 0 \tag{44}$$

Nous avons déjà exprimé que l'indice centré $[s - \mathcal{E}(S)]$, où $s = \text{card}[R(\omega) \cap R(\bar{\omega})]$, représente exactement dans ce cas, le numérateur du coefficient τ de M.G. Kendall. Dans ces conditions et dans le cas le plus général de la comparaison de deux préordres totaux ω et $\bar{\omega}$, nous sommes partis du même indice brut

$$s = card(A_1 \cap B_1) = card\left(\Sigma\left\{(E_i \cap F_j) \times (E_{i'} \cap F_{j'}) / 1 \le i \le i' \le h\,,\, 1 \le j \le j' \le k\right\}\right), \tag{45}$$

où nous notons ici $i, i', \ldots$ (resp. $j, j', \ldots$) des étiquettes de classes de ω (resp. $\bar{\omega}$). On a [Lerman 1973, 1981, 1983], sous l'hypothèse d'absence de liaison H_1 :

$$\mathcal{E}\,(S) = \lambda\mu \text{ et var } (S) = \lambda\mu + \rho_{cc}\sigma_{cc} + \rho_{ff}\sigma_{ff} + 2\rho_{cf}\sigma_{cf} + \theta\zeta - \lambda^2\mu^2.$$

$$\text{où } \lambda = \frac{1}{\sqrt{n(n-1)}} \sum \{m_i\, m_{i'} / 1 \le i < i' \le h\},$$

$$\rho_{cc} = \frac{1}{\sqrt{n(n-1)(n-2)}} \sum \{m_i\, m_{c(i)}[m_{c(i)} - 1] / 2 \le i \le h\}, \tag{46}$$

$$\rho_{ff} = \frac{1}{\sqrt{n(n-1)(n-2)}} \sum \{m_i\, m_{f(i)}[m_{f(i)} - 1] / 1 \le i \le (h-1)\},$$

$$\rho_{cf} = \frac{1}{\sqrt{n(n-1)(n-2)}} \sum \{m_i\, m_{c(i)}\, m_{f(i)} / 2 \le i \le (h-1)\}$$

$$\theta = \frac{1}{\sqrt{n(n-1)(n-2)(n-3)}} \times$$

$$\sum \left\{m_i\, m_{i'}\left[\sum \{m_p\, m_{p'} / 1 \le p < p' \le h\} + m_i + m_{i'} - 2n + 1\right] / 1 \le i < i' \le h\right\}$$

où on note $m_{c(i)} = \sum \{m_{i'} / i' < i\}$ et $m_{f(i)} = \sum \{m_i / i' > i\}$.

D'autre part, les expressions de μ, σ_{cc}, σ_{ff}, σ_{cf} et ζ sont respectivement de même forme que celles λ, ρ_{cc}, ρ_{ff}, ρ_{cf} et θ ; si les premières sont relatives à la composition $t(\omega)=(m_1,m_2,\ldots,m_h)$, les secondes sont relatives à la composition $t(\bar{\omega})=(n_1,n_2,\ldots,n_k)$.

On peut maintenant chercher à analyser la forme de l'indice centré [s - $\mathcal{E}$ (S)] par rapport aux paramètres (39). On a :

$$s - \mathcal{E}(S) = s(s + p + q + r) - (u + p)(v + q), \tag{47}$$

où le premier (resp. second) terme du second membre représente un aspect accord (resp. désaccord) entre les deux préordres totaux ω et $\overline{\omega}$.

N'étant plus dans un cas pouvant se ramener à un codage antisymétrique, la question se pose de reconnaître la forme limite dans les conditions de la propriété 2 ci-dessus, du coefficient suivant, que nous noterons $Q_1^0(\omega,\overline{\omega})$ pour le différencier de celui (41) ci-dessus :

$$Q_1^0(\omega,\overline{\omega}) = \frac{s - \mathcal{E}(S)}{\sqrt{var(S)}} \tag{48}$$

PROPRIÉTÉ 4. Sous les conditions de la propriété 2 ci-dessus, la forme limite du coefficient $Q_1^0(\omega,\overline{\omega})$ est $\frac{\sqrt{n}}{2} \times \frac{\nu}{\sqrt{\delta}}$, où :

$$\nu = \sum \left\{ (\nu_{ij}\nu_{i'j'} - \omega_i\ \omega_{i'}\overline{\omega}_j\ \overline{\omega}_{j'}) \,/\, 1 \le i < i' \le h\,,\ 1 \le j < j' \le k \right\} \tag{49}$$

et où :

$$\begin{aligned}
\delta = &\left(\sum_{i<i'} \omega_i\ \omega_{i'}\right)^2 \left(\sum_{j<j'} \overline{\omega}\ \overline{\omega}_{i'}\right)^2 \\
&\frac{1}{4}\left\{\left[\sum_i \omega_i\,(\omega_i^c)^2\right]\left[\sum_j \overline{\omega}_j\,(\overline{\omega}_j^c)^2\right] + \left[\sum_i \omega_i\,(\omega_i^f)^2\right]\left[\sum_j \overline{\omega}_j\,(\overline{\omega}_j^f)^2\right]\right\} \\
&+ \frac{1}{2}\left(\sum_i \omega_i\,\omega_{ij}^c\omega_i^f\right)\left(\sum_j \overline{\omega}_j\,\overline{\omega}_j^c\overline{\omega}_j^f\right) \\
&- \frac{1}{4}\left\{\left(\sum_{i<i'} \omega_i\ \omega_{i'}\right)^2\left[\sum_{j<j'} \overline{\omega}_i\ \overline{\omega}_{i'}(2 - \overline{\omega}_i - \overline{\omega}_{i'})\right] + \left(\sum_{j<j'} \overline{\omega}_j\,\overline{\omega}_j\right)^2\left[\sum_{i<i'} \omega_i\,\omega_{i'}\,(2 - \omega_i - \omega_{i'})\right]\right\},
\end{aligned} \tag{48}$$

où ω_i^c et ω_i^f (resp. $\overline{\omega}_j^c$ et ω_j^f sont les proportions correspondant aux fréquences absolues $m_{c(i)}$ et $m_{f(i)}$ [resp. $n_{c(j)}$ et $n_{f(j)}$].

Ce résultat prouve qu'on a le même phénomène quant à la tendance asymptotique du coefficient d'association $Q(\alpha,\beta)$ entre les deux relations binaires α et β sur O. Sous les conditions de la propriété 2 ci-dessus, la forme limite de $Q(\alpha,\beta)$ est $\sqrt{n}\ g\left[p\,(\alpha \wedge \beta)\right]$, où $g\left[p\,(\alpha \wedge \beta)\right]$ est une fonction du tableau $p(\alpha \wedge \beta)$ des proportions $\nu_{ij} = \frac{n_{ij}}{n}$, $1 \le i \le h$, $1 \le j \le k$.

6. CONCLUSION ET PERSPECTIVE

6.1. Travaux récents et en cours

Ces travaux s'organisent autour de deux thèses qui viennent de s'achever ; celles de F. Daudé et de M. Ouali-Allah. C'est ainsi que l'ensemble du calcul du paragraphe 4 a été repris avec une logique assez différente, ce qui a permis d'obtenir différentes expressions synthétiques et élégantes de la variance de l'indice brut aléatoire [Ouali-Allah 1991a]. D'autre part, suite à notre suggestion, M. Ouali-Allah (1991_a) a considéré la construction d'un coefficient d'association entre variables qualitatives "préordonnance". L'ensemble des couples de modalités d'une telle

variable est muni d'un préordre total, traduisant de façon ordinale les ressemblances orientées entre modalités. En codant ce préordre total au moyen de la fonction "rang moyen", on se ramène formellement à la comparaison de deux valuations sur $O \times O$. Mais, il faut se rendre compte que la base du calcul concernant un couple (v,w) de variables présentant respectivement h et k modalités, est une table de contingence de dimension h x k. Le codage en termes de préordonnance d'une variable descriptive, notamment qualitative, offre une grande souplesse ; d'où, l'élaboration d'un programme très général AVARE (Association entre VAriables RElationnelles) de calcul de la matrice des coefficients d'association sur un ensemble V de variables descriptives observées sur un ensemble O d'objets [M. Ouali-Allah 1991_b]. Un autre programme AVAND (Association entre VAriables à modalités Non Disjointes), traite du cas où, pour une entité donnée, une même variable descriptive est soit non concernée, soit détermine une distribution de probabilité.

Nous avons déjà exprimé (cf.§3) que suite à la question posée F. Daudé a étudié l'ensemble des trois modèles de l'hypothèse d'absence de liaison H_1, H_2 et H_3 pour la comparaison de deux variables qualitatives au moyen de la définition d'un tableau de contingence aléatoire [Daudé 1992] ; alors que, jusqu'à présent, seule H_1 avait été exploitée. Des résultats théoriques et de simulation à grande échelle permettent d'établir, dans la plupart des situations, le caractère asymptotiquement normal de la distribution de l'indice aléatoire. Mais, ce n'est pas toujours le cas ; surtout lorsqu'il s'agit d'associer deux variables qualitatives nominales à faible nombre de modalités et équidistribuées. Toutefois, des solutions de rechange peuvent être proposées pour les situations très caractéristiques où le modèle normal ne convient pas. Ces travaux sont suivis avec intérêt par Ph. Villoing (CCSA, Celar, Bruz) qui nous apporte une contribution logistique.

6.2. Généralisation au cas q-aire

Nous avons déjà signalé (cf.§1) que dans les situations concrètes rencontrées dans la pratique de l'analyse des données qualitatives, on recouvrait la totalité des cas en considérant q = 1,2,3 ou 4, où q est l'arité commune des deux relations à comparer. Jusqu'à présent nous nous sommes formellement ramenés aux cas q = 1 et q = 2. Nous avons commencé par mener les calculs concernant un q quelconque. Toutefois, pour un q donné, avec une méthode générale de calcul, il faut tenir compte de la spécificité commune des deux relations q-aires à comparer.

$O^{[q]}$ désignant l'ensemble des q-uples d'objets sans composante commune [card($O^{[q]}$) = $n(n - 1)...(n - q+1)$, où n = card(O)], nous venons de considérer le cas de la comparaison de deux valuations (pouvant être logiques; c'est-à-dire, binaires) sur $O^{[q]}$.

Les deux valuations peuvent être notées comme suit :

$$\left\{\mu_{i_1 i_2 \ldots i_q} / (i_1, i_2, \ldots, i_q) \in I^{[q]}\right\}, \tag{1}$$

$$\left\{\nu_{i_1 i_2 \ldots i_q} / (i_1, i_2, \ldots, i_q) \in I^{[q]}\right\}, \tag{2}$$

où I={1,2,...,i,...,n} indexe l'ensemble O des objets.

L'indice brut prend la forme suivante :

$$s(\mu,\nu) = \Sigma\left\{\mu_{i_1 i_2 \ldots i_q}\, \nu_{i_1 i_2 \ldots i_q} / (i_1, i_2, \ldots, i_q) \in I^{[q]}\right\} \tag{3}$$

μ^* et ν^* étant deux valuations aléatoires indépendantes conformément au modèle H_1 permutationnel de l'hypothèse d'absence de liaison, la distribution de $s(\mu^*,\nu^*)$ est la même que celle de $s(\mu,\nu^*)$ [resp. $s(\mu^*,\nu)$], où

$$s(\mu,\nu^*) = \Sigma \left\{ \mu_{i_1 i_2 \ldots i_q} \nu_{\sigma(i1)\sigma(i2)\ldots\sigma(iq)} / (i_1,i_2\ldots,i_q) \in I^{[q]} \right\} \tag{4}$$

On obtient par le calcul

$$\mathcal{E}^* = [s(\mu,\nu^*)] = n^{[q]} \bar{\mu}\, \bar{\nu}, \tag{5}$$

où on a noté $n^{[q]}$ pour $n(n-1) \times \ldots \times (n-q+1)$ et où $\bar{\mu}$(resp. $\bar{\nu}$) désigne la moyenne de la valuation μ (resp. ν) sur $O^{[q]}$.

Le calcul de la variance de $s(\mu,\nu^*)$ se ramène à celui du moment absolu d'ordre 2, $E[s^2(\mu,\nu^*)]$; et, on a :

$$s^2(\mu,\nu^*) = \sum_{I^{[q]} \times I^{[q]}} \mu_{i_1 i_2 \ldots i_q} \times \mu_{j_1 j_2 \ldots j_q} \times \nu_{\sigma(i_1)\sigma(i2)\ldots\sigma(i_q)} \times \nu_{\sigma(j_1)\sigma(j2)\ldots\sigma(j_q)} \tag{6}$$

Le problème concerne l'évaluation de

$$\mathcal{E}^* \left(\nu_{\sigma(i_1)\sigma(i2)\ldots\sigma(i_q)} \times \nu_{\sigma(j_1)\sigma(j2)\ldots\sigma(j_q)} \right) \tag{7}$$

Pour $(i_1,i_2,\ldots,i_q)$ donné, on a l'invariance conditionnelle de cette espérance mathématique dès lors que $(j_1,j_2,\ldots,j_q)$ a, par rapport à $(i_1,i_2,\ldots,i_q)$, une configuration fixée. Une telle configuration relative se trouve définie selon que des composantes du premier q-uple se répètent en certaines positions du second q-uple. Ainsi, si ν est le nombre de composantes du premier q-uple qui se répètent dans le second q-uple, il y a, en notant $\binom{m}{l}$ le coefficient binomial $m!/l!(m-l)!$,

$$\binom{q}{r}^2 r\,!\,, \tag{8}$$

Ainsi, le nombre total de configurations est

$$\sum_{0 \le r \le q} \binom{q}{r}^2 r\,!\,. \tag{9}$$

La complexité du calcul que supporte ce nombre est parfaitement abordable pour q petit ; ainsi, pour $q = 4$ (resp. $q = 5$), ce nombre vaut 209 (resp. 1546).

Une même configuration qui suppose la répétition de r composantes de $(i_1,i_2,\ldots,i_q)$, se trouve représentée par $\binom{n-q}{q-r}(q-r)\,!$ points de $I^{[q]}$ et d'ailleurs, on peut vérifier que

$$\sum_{0 \le r \le q} \binom{q}{r}^2 r\,! \binom{n-q}{q-r}(q-r)\,! = n^{[q]}. \tag{10}$$

Dans ces conditions, pour évaluer $\mathcal{E}^*[s^2(\mu,\nu^*)]$, on décomposera l'ensemble $I^{[q]} \times I^{[q]}$ d'indexation de la somme selon les différentes configurations du couple des q-uples $[(i_1,i_2,\ldots,i_q),(j_1,j_2,\ldots,j_q)]$. Si c_r est l'une des configurations où r composantes du premier q-uple se répètent dans le second, la partie de $\mathcal{E}^*(s^2(\mu,\nu^*)]$ relative à l'ensemble $C(c_r)$ des couples de q-uples de même configuration c_r est égal à

$$\frac{1}{n[2q+r]}\left(\sum_{C_{(c,r)}} \mu_{i_1 i_2 \ldots i_q} \; \mu_{j_1 j_2 \ldots j_q}\right)\left(\sum_{C_{(c,r)}} \nu_{i_1 \ldots i_q} \nu_{j_1 j_2 \ldots j_q}\right) \qquad (11)$$

Ce qui, avec bien sûr l'aide d'un programme informatique, résout le problème du calcul de la variance de $s(\mu,\nu^*)$ et donc celui, du coefficient d'association statistiquement normalisé $Q_1(\mu,\nu)$, conforme au schéma de la figure 1 (§1), entre les deux valuations q-aires μ et ν;

Grâce à la démarche sous jacente à cet outil général nous envisageons différentes études dans des cas plus particuliers et plus typés ; il s'agit par exemple de l'élaboration d'un coefficient d'association Q_1 entre deux préordonnances totales ; mais en situant cette fois ci la représentation au niveau de $O^{[2]} \times O^{[2]}$ et non plus de $O^{[2]}$ — à partir de la notion de "rang moyen" — comme cela a été mentionné ci-dessus [cf.§6.1].

6.3. Impact et situation de cette recherche

Le problème que pose l'élaboration de l'expression formelle d'un coefficient d'association entre variables relationnelles et notamment qualitatives, est considérable et occupe une part importante de l'analyse combinatoire des données [Arabie & Hubert 1992]. Dans une telle construction, il est très difficile de ne pas subir l'arbitraire du choix. Notre approche dont nous avons pu montré l'extrême généralité permet de réduire considérablement cet arbitraire en le localisant au niveau du choix de l'indice brut où alors, il peut être beaucoup plus facilement paramétré. Ainsi, il s'agit d'associer deux variables qualitatives ordinales, on peut considérer avec les notations de (39) du précédent paragraphe et sans pour ainsi dire d'arbitraire, un indice brut de la forme

$$i(\omega,\overline{\omega}) = s+\alpha\, r - u, \qquad (12)$$

où α est un paramètre positif compris entre 0 et 1.

On aura d'autre part à choisir une forme de l'hypothèse d'absence de liaison. Il s'agira de H_1 ou de H_3 qui s'en distingue de façon perceptible, non au niveau du centrage de l'indice brut ; mais, à celui de la réduction.

Le coefficient normalisé a une propriété formelle d'invariance par rapport à toute transformation affine ne dépendant que des distributions marginales empiriques des deux variables à comparer. D'autre part, compte tenu de la manière dont il a été obtenu, notre coefficient met en évidence la signification statistique de l'association des deux relations à comparer. D'ailleurs, cette approche a un solide ancrage puisque nous retrouvons des coefficients aussi bien établis que ceux de K. Pearson, Bravais-Pearson, M.G. Kendall,...

Enfin, l'expérience pratique montre que l'arbitraire du choix à niveau de l'indice brut, se trouve neutralisé de façon appréciable par la normalisation statistique que nous considérons. D'ailleurs, nous avons pu nous rendre compte [cf. expression (47) §5] qu'en partant -dans le cas qualitatif ordinal- de l'indice brut s, on aboutissait après centrage à la différence entre deux expressions ; dont la première marque l'accord et la seconde, le désaccord entre les deux préordres totaux ω et $\overline{\omega}$. L'influence du choix de l'indice brut est d'autant plus neutralisée s'il s'agit de comparer de façon seulement ordinale les associations entre paires de variables.

Nous aurions pu nous poser la question préalable : mais enfin, pourquoi un coefficient d'association entre variables qualitatives ?

Si notre seul univers se limitait à l'observation de deux variables v et w sur un ensemble O d'objets à un instant donné. L'intérêt de la construction et de l'évaluation d'un coefficient

d'association $c(v,w)$ est négligeable. Un certain intérêt peut être lié à l'évolution d'un tel coefficient lorsque l'ensemble O varie ; il peut s'agir d'une évolution dans le temps, comme il peut s'agir de la situation statistique déjà mentionnée au paragraphe 1 où O est un échantillon aléatoire de taille croissante d'une population $\mathfrak{P}$.

Nous situons quant à nous l'intérêt le plus vif au niveau de la comparaison deux à deux d'un ensemble A de variables descriptives, observées sur un ensemble O d'objets ou d'ailleurs (cf. ci-dessous) sur un ensemble C de classes. Plus précisément, c'est par rapport à l'objectif de l'organisation selon un arbre de classification, des liens mutuels entre attributs de A, que nous avons mené notre recherche sur les coefficients d'association entre variables descriptives. Nous avons une méthode pour détecter les noeuds "significatifs" ou "pertinents" d'un arbre de classification [Lerman & Ghazzali 1991]. Dans ces conditions, la décomposition de A en classes et sous-classes significatives d'association permet la découverte d'un système hiérarchique de "facteurs" et sous "facteurs" ; les sous facteurs étant relativement indépendants à l'intérieur d'un même facteur plus général. Par conséquent, il s'agit d'un terme très intéressant à l'alternative posée par l'analyse factorielle linéaire des données qui doit supposer une représentation géométrique des varaibles descriptives.

Ainsi donc, l'élaboration de coefficients d'association entre variables descriptives, n'a pas pour nous, comme d'ailleurs pour toute l'analyse des données, comme but de tester les mutuelles indépendances entre variables. En effet, comme nous l'avons montré dans [Lerman 1984], la philosophie de l'analyse des données est en quelque sorte opposée à celle des tests d'indépendance. Pour cette dernière, on a, relativement à l'existence d'un lien :

FAUX JUSQU'A PREUVE DU CONTRAIRE ;

alors que pour l'optique des données on a :

VRAI JUSQU'A PREUVE DU CONTRAIRE.

Dans ces conditions, il y a alors lieu d'évaluer et d'organiser au mieux les liens mutuels. A cette fin, le travail mené ici correspond certes à une étape très importante, mais non définitive. En effet, l'association entre deux variables données n'a pas un caractère relatif par rapport au contexte de l'ensemble des liaisons deux à deux observées sur l'ensemble A des variables. En fait, nous tenons compte de ce contexte pour aboutir à une échelle de probabilité pour l'évaluation des liens que nous organisons au moyen de l'algorithmique de la construction ascendante hiérarchique d'un arbre de classification. Une telle construction est dans notre cas fondée sur la famille de critères de la vraisemblance du lien maximal [Lerman 1991].

La structure de la donnée à laquelle nous nous sommes référés le long de toute cette étude correspond à ce qu'on appelle un système de Tarski [Tarski 1954], dont la forme générale est :

$$T = <O;R_1,R_2,...,R_m> \qquad (13)$$

où $R_1,R_2,...,R_{m-1}$ et R_m sont m relations définies sur l'ensemble O des objets et que nous supposons de même arité q.

Un autre type de système que nous avons pu faire émerger se présente comme suit :

$$S = <C;R_1,R_2,...,R_m> \qquad (14)$$

Ici C est un ensemble de classes (on dit encore concepts, surtout lorsque les classes sont définies en intention). $R_1,R_2,...,R_{m-1}$ et R_m sont comme ci-dessus des relations de même arité q. Cependant, ce dont on dispose maintenant, c'est de la distribution statistique de chacune des relations $R_j (1 \leq j \leq m)$ sur chacune des classes c faisant partie de C.

Nous avons été conduits à l'introduction du système S pour représenter fidèlement et traiter par la classification une base de connaissance très structurée du point de vue de l'intelligence artificielle. Signalons qu'un cas très particulier de S correspond à une juxtaposition horizontale de tableaux de contingence. On comprend dans ces conditions l'extension nécessaire de nos coefficients d'association pour les comparaisons mutuelles entre les relations $R_j, 1 \leq j \leq m$, sur la base du système S. D'importants résultats, dans des situations spécifiques, ont déjà été obtenus et sont en cours d'expérimentation. Il s'agit notamment du cas où chaque R_j est une relation préordonnance.

C'est surtout le problème de la classification d'un ensemble O d'objets, éventuellement pondéré et muni d'un indice de distance ou de dissimilarité, qui a occupé la vision classique de la littérature taxinomique. Dans une telle vision, le cas d'importance est celui où l'indice de distance résulte de la représentation de l'ensemble O d'objets par un nuage de points dans un espace euclidien dont chaque axe est associé à une variable de description; cette dernière étant interprétée comme une projection sur l'axe concerné. Nous considérons également ce problème de la classification d'un ensemble O d'objets pondéré, sur la base d'une donnée telle que (13) ci-dessus qui -bien sûr, comme nous l'avons déjà exprimé- comprend comme cas particulier celui où chaque $R_j, 1 \leq j \leq m$, est associée à une variable quantitative, qui est ainsi interprétée comme une valuation de O. Nous étudions d'autre part, le problème de la classification d'un C de classes sur la base d'un système tel que S [Lerman & Peter 1989, Lerman 1991].

Une dernière étape d'un système d'analyse des données par la classification concerne la mise en correspondance d'un système organisé de classes sur l'ensemble A des attributs avec un système organisé de classes sur l'ensemble O des objets à partir de la donnée du système T [cf. (13)] (resp. sur l'ensemble C des concepts à partir de la donnée du système S [cf. (14)]). Les idées développées dans ce travail nous permettent de le faire ; comme en attestent de nombreux résultats déjà obtenus.

BIBLIOGRAPHIE

ARABIE P. and HUBERT L.J. (1992), "Combinatorial data analysis", 1992, *Annual Review of Psychology*, 43, pp. 169-203.

CHAH S. (1984), "Agrégation des préordonnances", *Etude F-063*, Centre Scientifique IBM de Paris.

CHAH S (1985) "Critères de classification sur des données hétérogènes", *Proceedings of the fourth international symposium on data analysis and informatics*, edited by E. Diday and al, North Holland, 1986.

DANIELS H.E. (1944), "The relation between measures of correlation in the universe of sample permutations", *Biometrika*, vol. 33, 129-135.

DAUDE F., "Normalisation sous hypothèses d'absence de lien", *Publication interne IRISA*, Rennes, n° 549, Sept. 1990, 42 pages.

DAUDE F. (1992), *Analyse et justification de la notion de ressemblance entre variables qualitatives dans l'optique de la classification hiérarchique par AVL*, Thèse de l'Université de Rennes I, 24 Juin 1992 (à paraître).

EFRON B. (1986), "The Jacknife, the Boot-strap and other resampling plans", *CBMS-NSF regional conference series in applied mathematics*.

GIAKOUMAKIS V. et MONJARDET B. (1987), "Coefficients d'accord entre deux préordres totaux", *Statistique et Analyse des Données* 12, pp. 46-99.

GOODMAN L.A. and KRUSKAL W.H. (1954), "Measures of association for cross classifications", *Journal of the American Statistical Association*, Vol. 49, pp. 732-764.

GOODMAN L.A. and KRUSKAL W.H. (1963), "Measures of association for cross classifications" III : Approximate sampling theory",*Journal of the American Statistical Association* Vol. 58, pp. 310-364.
HAJEK J. (1961), "Some extensions of the Wald-Wolfowitz-Noether theorem", *AMS*, 32, pp. 506-523.
HUBERT L.J. (1983), "Inference procedures for the evaluation and comparaison of proximity matrices", *Numerical Taxonomy*, Ed. J. Felsenstein, NATO ASI Series, Berlin, Springer Verlag.
HUBERT L.J. (1987), *Assignment methods in combinatorial data analysis*, New York, Marcel Decker.
KENDALL M.G. (1970), *Rank correlation methods*, London, Charles Griffin, fourth edition (first edition in 1948).
LECALVE G. (1976), "Un indice de similarité pour des variables de types quelconques", *Statistique et Analyse des Données*, 01-02, pp. 39-47.
LERMAN I.C. (1973), "Etude distributionnelle de statistiques de proximité entre structures finies de même type ; application à la classification automatique", *Cahiers du Buro*, n° 19, Paris.
LERMAN I.C. (1976), "Formal analysis of a general notion of proximity between variables". *Congrès Européen des Statisticiens*, Grenoble, Sept. 1976, North Holland (1977).
LERMAN I.C. (1981), *Classification et analyse ordinale des données*, Paris, Dunod.
LERMAN I.C. (1983), "Association entre variables qualitatives ordinales nettes ou floues", *Statistique et Analyse des Données*, vol. 8 n° 7, pp. 41-73.
LERMAN I.C. (1984), "Justification et validité statistique d'une échelle [0,1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées", *Publ. Inst. Stat. Univ. Paris*, XXIX, fasc. 3-4, pp. 27-57.
LERMAN I.C. (1987_a), "Construction d'un indice de similarité entre objets décrits par des variables d'un type quelconque. Application au problème du consensus en classification", *Rev. Statistique Appliquée*, XXXV (2), pp. 39-60.
LERMAN I.C. (1987_b), "Analyse de la forme limite de coefficients statistiques d'association entre variables relationnelles", *Rapport de recherche n° 702*, Inria, Juillet 1987.
LERMAN I.C. (1987_c), "Maximisation de l'association entre deux variables qualitatives ordinales", *Math. Sci. hum.* 25ème année, n° 100, 1987, pp. 49-56.
LERMAN I.C. (1988), "Structure maximale pour la somme des carrés d'une contingence aux marges fixées; une solution algorithmique programmée", *Rairo*, vol. 22, n° 2, pp. 83 à 136.
LERMAN I.C. (1991), "Foundations of the Likelihood Linkage Analysis (LLA) Classification method", *Applied Stochastic Models and Data Analysis*, vol. 7, pp. 63-76 (J. Wiley).
LERMAN I.C. et GHAZZALI N. (1991), "Quoi retenir d'un arbre de classification ? Un essai en quantification d'image numérisée", *Rapport de recherche n° 1386*, Inria, Janvier 1991.
LERMAN I.C., GRAS R. et ROSTAM H. (1981), "Elaboration et évaluation d'un indice d'implication pour des données binaires" I et II ; I : *Math. & Sci. hum.*, 19ème année, n° 74, 1981 pp. 5-35, II : *Math. & Sci. hum.*, 19ème année, n° 75, 1981, pp. 5-47.
LERMAN I.C. et PETER Ph. (1985), "Organisation et consultation d'une banque de "petites annonces" à partir d'une méthode de classification hiérarchique en parallèle", *Journées Internationales Analyse des Données et Informatique IV, Octobre 1985, Versailles*, North Holland (1986), pp. 121-136.
LERMAN I.C. et PETER Ph. (1989), "Classification of concepts described by taxonomic preordonnance variables with multiple choice. Application to the structuration of a species set of phebotomine" *Data Analysis, Learning symbolic and numerical knowledge*, edited by E. Diday, Inria, Nova Science Publishers, (1989), pp. 73-87.
MANTEL N. (1967), "Detection of disease clustering and a generalized regression approach", *Cancer Research*, vol. 27, n° 2, pp. 209-220.
MESSATFA H. (1990), *Unification relationnelle des critères et structures optimales des tables de contingences*, thèse de doctorat de l'Université de Paris 6, 5 mars 1990.

MIELKE W. (1979), "On asymptotic non normality of null distributions of MRPP Statistics", *Communications in Statistics, Theory and Methods*, A8 (15), pp. 1541-1550.
NOETHER G. (1949), "On a theorem by Wald and Wolfowitz", *Ann. Math. Stat.* vol. 20, pp. 455-458.
OUALI-ALLAH M. (1991_a), *Analyse en préordonnances des données qualitatives. Applications aux données numériques et symboliques*, Thèse de l'Université de Rennes I, 5 décembre 1991.
OUALI-ALLAH M. (1991_b), "Avare : un programme de calcul des associations entre variables relationnelles", *Publication Interne Irisa* n° 591, juin 1991, 32 pages.
PETER Ph. (1987), *Méthodes de classification hiérarchique et problèmes de structuration et de recherche d'informations, assistées par ordinateur*, thèse de l'Université de Rennes I, 6 mars 1987.
SUPPES P. and ZINNES J.L. (1963), "Basic measurement theory" *Handbook of mathematical psychology*, Eds Bush, Luce, Galanter, New York, J. Wiley, pp. 2-76.
TARSKI A. (1954), "Contribution to the theory of models", I.II. *Indagationes Mathematicae*, 16, pp. 572-588.
WALD A. and WOLFOWITZ J. (1944), "Statistical tests based on permutations of the observations", *Ann. Math. Stat.* vol. 15, pp. 358-372.