Analyse de la forme limite de coefficients statistique d'association entre variables relationnelles

Israël-César Lerman

INRIA

UNITÉ DE RECHERCHE
INRIA-RENNES

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
BP 105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Rapports de Recherche

N° 702

ANALYSE DE LA FORME LIMITE DE COEFFICIENTS STATISTIQUE D'ASSOCIATION ENTRE VARIABLES RELATIONNELLES

Israel Cesar LERMAN

JUILLET 1987

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTÈMES ALÉATOIRES

Campus Universitaire de Beaulieu
35042 - RENNES CÉDEX
FRANCE
Téléphone : 99 36 20 00
Télex : UNIRISA 950 473 F
Télécopie : 99 38 38 32

ANALYSE DE LA FORME LIMITE DE COEFFICIENTS STATISTIQUE D'ASSOCIATION ENTRE VARIABLES RELATIONNELLES

Israël-César **LERMAN**

Publication Interne n° 367 - Juin 1987 - 58 Pages

ANALYSIS OF THE LIMIT FORM OF STATISTICAL ASSOCIATION COEFFICIENTS BETWEEN RELATIONAL VARIABLES

Israël-César **LERMAN**

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
(L. A. 227)
UNIVERSITÉ DE RENNES 1 — I. N. S. A. DE RENNES

INSTITUT NATIONAL DE RECHERCHE
EN INFORMATIQUE ET EN AUTOMATIQUE
(LABORATOIRE DE RENNES)

ANALYSE DE LA FORME LIMITE DE COEFFICIENTS STATISTIQUE D'ASSOCIATION ENTRE VARIABLES RELATIONNELLES

Israël César LERMAN

RESUME :

On considère un coefficient général d'association entre deux variables relationnelles dont la normalisation résulte de la variance d'un indice aléatoire ayant une distribution permutationnelle. On établit les liens entre différentes expressions de la variance et on détermine deux nouvelles expressions dont l'une permet l'étude de la forme limite du coefficient d'association dans des conditions qu'on appréhende clairement. On considère les cas 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 à partir d'une démarche combinatoire et statistique.

ANALYSIS OF THE LIMIT FORM OF STATISTICAL ASSOCIATION COEFFICIENTS BETWEEN RELATIONAL VARIABLES.

ABSTRACT

We consider a general association coefficient between two relational variables, the normalization of which being obtained by the means of the variance of a random index having permutational distribution. We establish correspondence between different expressions of the variance and we determine two new expressions. One of these last enables us to study – under clear conditions – the limit form of the association coefficient. More particulary, we consider the important cases of the comparison between two nominal or ordinal qualitative variables. The limit expression makes possible to realize from purely formal point of view, the nature of the normalization obtained by combinatorial and statistical approach.

I_ COMPARAISON STATISTIQUE DE DEUX VARIABLES RELATIONNELLES . RAPPEL ET OBJECTIF DE L'ETUDE.

Pour comparer deux variables qualitatives relationnelles, nous avons au cours de notre recherche fait émerger le diagramme suivant :

$$(\alpha, \beta) \in A \times B \longrightarrow [R(\alpha), R(\beta)] \in \Omega_\alpha \times \Omega_\beta \quad (1)$$

$$\longrightarrow \quad s = s(\alpha, \beta) = \text{card}\,[R(\alpha) \cap R(\beta)] \quad (2)$$

→ Hypothèse d'absence de lien (h.a.l.) (ou d'indépendance) tenant en compte de façon stricte ou 'floue' les caractéristiques de cardinalité de α et de β. (3)

$$\longrightarrow \quad S = s(\alpha^*, \beta^*) = \text{card}\,[R(\alpha^*) \cap R(\beta^*)] \quad (4)$$

$$\longrightarrow \quad Q(\alpha, \beta) = [s - \mathcal{E}(S)] / \sqrt{\text{var}(S)} \quad (5)$$

Dans ce schéma α et β sont les deux relations sur l'ensemble O des objets, respectivement déterminées par les deux variables à comparer. A (resp. B) est l'ensemble de toutes les relations de « même type » que α (resp. β). $R(\alpha)$ [resp. $R(\beta)$] est la représentation ensembliste de α [resp. β]. $R(\alpha)$ [resp. $R(\beta)$] est un sous ensemble de O ou bien de $O \times O$, ou même de $(O \times O) \times (O \times O)$. Ω_α (resp. Ω_β) est l'ensemble de tous les sous ensembles possibles de représentation d'une relation de même type que α (resp. β). $S = S(\alpha, \beta)$ est appelé indice "brut". α^* et β^* sont deux variables relationnelles aléatoires indépendantes, respectivement associées à α et à β, conformément à l'hypothèse d'absence de liaison, qui tient compte – de façon stricte ou floue – des caractéristiques cardinales de α et de β. S est l' "indice brut aléatoire" dont l'espérance mathématique et la variance sont

notées $\mathcal{E}(S)$ et $var(S)$. $Q(\alpha, \beta)$ est l'indice "centré et réduit".

Pour clarifier ce shéma, nous allons l'illustrer dans deux situations qui se présentent classiquement en analyse des données ; la première est celle où on a à comparer deux variables qualitatives nominales et la seconde est celle où on a à comparer deux variables qualitatives totalement ordinales.

Dans la première situation α et β sont deux partitions dont on supposera — sans restreindre la généralité — qu'elles sont en classes étiquetées. Nous les désignerons-pour spécifier les expressions ci-dessous — par π et x. $t(\pi)$ [resp. $t(x)$] indiquera le type de la partition π (resp. x) ; c'est à dire, la suite ordonnée des cardinaux

de ses classes. Plus précisément, en posant $\pi = (E_i / 1 \leqslant i \leqslant h)$, $t(\pi) = (m_i / 1 \leqslant i \leqslant h)$, où $m_i = \mathrm{card}(E_i)$ et où $\sum \{m_i / 1 \leqslant i \leqslant h\} = n = \mathrm{card}(O)$. De même, en posant $\chi = (F_j / 1 \leqslant j \leqslant k)$, $t(\chi) = (n_j / 1 \leqslant j \leqslant k)$, où $n_j = \mathrm{card}(F_j)$ et où $\sum \{n_j / 1 \leqslant j \leqslant k\} = n$. Dans ces conditions A (resp. B) est l'ensemble des partitions - en classes étiquetées - sur O, de type $t(\pi)$ [resp. $t(\chi)$]. Introduisons l'ensemble $O^{\{2\}} = P_2(O)$ des parties à deux éléments de O. $R(\pi)$ [resp. $R(\chi)$] est l'ensemble des paires d'objets dont les deux composantes sont réunies dans une même classe de la partition π (resp. χ). Plus précisément, $R(\pi) = \sum \{E_i^{\{2\}} / 1 \leqslant i \leqslant h\}$ [resp. $R(\chi) = \sum \{F_j^{\{2\}} / 1 \leqslant j \leqslant k\}$]. Ω_π (resp. Ω_χ) peut 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. $t(\chi)$].

$$s = s(\pi, \chi) = \text{card}\,[R(\pi) \cap R(\chi)] = \text{card}\,[R(\pi \wedge \chi)]$$

$$= \text{card}\left[\bigcup \{ (E_i \cap F_j)^{[2]} / 1 \leqslant i \leqslant h,\ 1 \leqslant j \leqslant k \}\right]$$

$$= \bigcup \{ n_{ij}(n_{ij}-1)/2 \,/\, 1 \leqslant i \leqslant h,\ 1 \leqslant j \leqslant k \}, \qquad (6)$$

où nous avons noté $n_{ij} = \text{card}\,(E_i \cap F_j)$, $1 \leqslant i \leqslant h$, $1 \leqslant j \leqslant k$.

Il y a trois formes fondamentales de l'h.a.l. [Lerman (1981) Chap. 2]. Nous allons considérer ici celle stricte où π^* (resp. χ^*) est une partition aléatoire dans l'ensemble A (resp. B) muni d'une probabilité uniformément répartie. La moyenne et la variance de l'indice brut aléatoire $S = s(\pi^*, \chi^*)$ sont respectivement donnés par [Lerman (1973), (1981)] :

$$\mathcal{E}(S) = \lambda\mu \quad \text{et} \quad \mathrm{var}(S) = \lambda\mu + \rho\sigma + \theta\varsigma - \lambda^2\mu^2,$$

où $\lambda = \sum_i \{ m_i(m_i-1)/\sqrt{2n(n-1)} \,/\, 1 \leqslant i \leqslant h \}$, (7)

$$\rho = \sum_i \{ m_i(m_i-1)(m_i-2)/\sqrt{n(n-1)(n-2)} \,/\, 1 \leqslant i \leqslant h \},$$

$$\theta = \{ [\sum_i m_i(m_i-1)]^2 - 2\sum_i m_i(m_i-1)(2m_i-3) \} / 2\sqrt{n(n-1)(n-2)(n-3)}$$

et où les expressions de μ, σ et ς ont respectivement la même forme que λ, ρ et θ ; les m_i de $t(\pi)$ étant remplacés par les n_j de $t(\chi)$.

On remarquera que θ peut s'exprimer en fonction de λ et de ρ puisque

$$\sum_i m_i(m_i-1)(2m_i-3) = 2\sum_i m_i(m_i-1)(m_i-2) + \sum_i m_i(m_i-1). \quad (8)$$

Dans la deuxième situation α et β sont deux préordres

totaux que nous noterons ω et ϖ. $t(\omega)$ [resp. $t(\varpi)$] indiquera la 'composition' de ω (resp. ϖ) ; c'est à dire, la suite des cardinaux de la suite ordonnée de ses classes. On adoptera des notations analogues à ci-dessus et on désignera par $(E_i / 1 \leqslant i \leqslant h)$ [resp. $(F_j / 1 \leqslant j \leqslant k)$] la suite ordonnée des classes de ω (resp. ϖ). $m_i = \mathrm{card}(E_i)$ et $n_j = \mathrm{card}(F_j)$, $1 \leqslant i \leqslant h$, $1 \leqslant j \leqslant k$. Nous introduisons ici l'ensemble $O^{[2]}$ des couples d'objets distincts de O. $R(\omega)$ [resp. $R(\varpi)$] est l'ensemble des couples d'objets (x, y) de $O^{[2]}$ tels que x précède strictement y pour ω (resp. ϖ). Plus précisément, $R(\omega) = \bigcup \{ E_i \times E_{i'} / 1 \leqslant i < i' \leqslant h \}$ [resp. $R(\varpi) = \bigcup \{ F_j \times F_{j'} / 1 \leqslant j < j' \leqslant k \}$]. Ω_ω (resp. Ω_ϖ) peut être défini comme étant l'ensemble des parties de $O^{[2]}$ dont chacune corres-

pond à la représentation d'un préordre total de composition $t(\omega)$ [resp. $t(\varpi)$] .

$$s = s(\omega, \varpi) = \text{card}\,[R(\omega) \cap R(\varpi)] = \text{card} \bigcup \left\{ (E_i \cap F_j) \times (E_{i'} \cap F_{j'}) \,/\, 1 \leqslant i < i' \leqslant h,\ 1 \leqslant j < j' \leqslant k \right\}$$

$$= \sum \left\{ n_{ij} \times n_{i'j'} \,/\, 1 \leqslant i < i' \leqslant h,\ 1 \leqslant j < j' \leqslant k \right\}, \qquad (9)$$

où nous notons $n_{ij} = \text{card}\,(E_i \cap F_j)$, $1 \leqslant i \leqslant h$, $1 \leqslant j \leqslant k$.

L'hypothèse d'absence de liaison associe à ω (resp. ϖ) un préordre aléatoire ω^* (resp. ϖ^*) dans l'ensemble - muni d'une probabilité uniforme - A (resp. B) , où A (resp. B) est l'ensemble des préordres totaux sur O de même composition $t(\omega)$ [resp. $t(\varpi)$] . La moyenne et la variance de l'indice brut aléatoire $S = s(\omega^*, \varpi^*)$ sont respectivement donnés par [Lerman (1973), (1981),

(1983)] :

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

où $\lambda = \dfrac{1}{\sqrt{n(n-1)}} \sum \{ m_i m_{i'} / 1 \leqslant i < i' \leqslant h \}$, (10)

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

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

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

et

$$\theta = \frac{1}{\sqrt{n(n-1)(n-2)(n-3)}} \sum \Big\{ m_i m_{i'} \Big[\sum \{ m_h m_{h'} / 1 \leqslant h < h' \leqslant h \} + m_i + m_{i'} - 2n + 1 \Big] / 1 \leqslant i < i' \leqslant h \Big\}$$

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 ς sont respectivement

de même forme que celles λ, ρ_{cc}, ρ_{cf} et θ ; si les premières sont relatives à la composition $t(\omega)$, les secondes sont relatives à la composition $t(\varpi)$.

On remarquera que θ peut s'exprimer en fonction des autres paramètres puisque

$$(m_i + m_{i'} - 2n + 1) = -[(n - m_i) + (n - m_{i'}) - 1]$$
$$= -[m_{c(i)} + m_{f(i)} + m_{c(i')} + m_{f(i')} - 1]. \quad (11)$$

Le schéma précédent (cf. début du paragraphe) constitue un puissant outil de conception de coefficients de comparaison entre variables qualitatives relationnelles. On retrouve comme cas particulier un indice aussi classique que celui de K. Pearson. Nous l'avons très extensivement utilisé pour construire nos indices d'association totale ou partielle [Lerman (1981), (1983a), (1983b)]. L'analyse conformément à ce schéma permet la mise en évidence de tous les phénomènes combinatoires et de calcul dans la

comparaison. La généralisation de cette approche à la comparaison de deux 'codages' ou 'pondérations' quelconques de $O^{[2]}$ est la suivante.

Soit ici $I=\{1,2,\dots,i,\dots,n\}$ l'ensemble indéxant l'ensemble des objets. $I^{[2]}=\{(i,j)/\ 1\leqslant i\neq j\leqslant n\}$ est l'ensemble des couples d'indices distincts. Soient $x=\{x_{ij}/(i,j)\in I^{[2]}\}$ et $y=\{y_{ij}/(i,j)\in I^{[2]}\}$ les deux 'codages' ou 'pondérations' de $O^{[2]}$. L'indice 'brut' s prend la forme suivante :

$$s=s(x,y)=\sum\{x_{ij}y_{ij}/(i,j)\in I^{[2]}\}\ . \quad (12)$$

Si on prend une forme bilatérale de l'h.a.l., l'indice brut aléatoire prend la forme suivante

$$S=s(x^*,y^*)=\sum\{x_{\sigma(i)\sigma(j)}y_{\tau(i)\tau(j)}/(i,j)\in I^{[2]}\}, \quad (13)$$

où σ et τ sont deux permutations aléatoires indépendantes prises dans l'ensemble G_n – muni d'une probabilité uniforme – des $n!$ permutations sur I.

S'inspirant d'un vieux papier de H.E. Daniels [Daniels (1944)], G. Lecalvé [Lecalvé (1976)] a eu – dans notre contexte – l'idée d'une telle extension. Nous avons repris cette étude de façon plus précisément combinatoire [Lerman (1976) repris en (1981)] ce qui nous a conduit à une expression formelle claire des moments de la distribution de $S = s(x^*, y^*)$. Les expressions de la moyenne $\mathcal{E}(S)$ et de la variance $\mathrm{var}(S)$ nécessitent l'introduction des ensembles suivants d'indéxation où des lettres différentes indiquent des indices distincts:

$$I^{[2]}, \quad D = \{[(i,j),(i,j)]\}, \quad E = \{[(i,j),(j,i)]\},$$

$$G_1 = \{[(i,j),(i,k)]\}\ ,\ G_1' = \{[(i,j),(h,i)]\},$$

$$G_2 = \{[(i,j),(h,j)]\}\ ,\ G_2' = \{[(i,j),(j,k)]\},$$

$$H = \{[(i,j),(h,k)]\}.$$

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

$$\mathcal{E}(S) = \frac{1}{n^{[2]}}\Big(\sum\{x_{ij}/(i,j)\in I^{[2]}\}\Big)\Big(\sum\{y_{ij}/(i,j)\in I^{[2]}\}\Big)$$

$$\operatorname{var}(S) = \frac{1}{n^{[2]}}\Big(\sum_{I^{[2]}} x_{ij}^2\Big)\Big(\sum_{I^{[2]}} y_{ij}^2\Big)$$

$$+ \frac{1}{n^{[2]}}\Big(\sum_{I^{[2]}} x_{ij}x_{ji}\Big)\Big(\sum_{I^{[2]}} y_{ij}y_{ji}\Big)$$

$$+ \frac{1}{n^{[3]}}\Big(\sum_{G_1} x_{ij}x_{ik}\Big)\Big(\sum_{G_1} y_{ij}y_{ik}\Big)$$

$$+\frac{1}{n^{[3]}}\Big(\sum_{G_1'} x_{ij}x_{hi}\Big)\Big(\sum_{G_1'} y_{ij}y_{hi}\Big)$$

$$+\frac{1}{n^{[3]}}\Big(\sum_{G_2} x_{ij}x_{hj}\Big)\Big(\sum_{G_2} y_{ij}y_{hj}\Big)$$

$$+\frac{1}{n^{[3]}}\Big(\sum_{G_2'} x_{ij}x_{jk}\Big)\Big(\sum_{G_2'} y_{ij}y_{jk}\Big)$$

$$+\frac{1}{n^{[4]}}\Big(\sum_{H} x_{ij}x_{hk}\Big)\Big(\sum_{H} y_{ij}y_{hk}\Big)$$

$$-\Big[\frac{1}{n^{[2]}}\Big(\sum_{I^{[2]}} x_{ij}\Big)\Big(\sum_{I^{[2]}} y_{ij}\Big)\Big]^2 . \qquad (14)$$

Lors de notre étude nous ignorions la contribution de N. Mantel [Mantel (1967)] qui - dans une optique de régression - considère la même statistique S. Dans ce papier qui ne mentionne toutefois pas la tentative de H.E. Daniels, on

trouve de façon claire une expression de la variance de S. Pour la présenter, introduisons les paramètres suivants :

$$A_1 = \Big(\sum_{I^{[2]}} x_{ij}\Big)^2 \ , \ A_2 = \sum_{i \in I} \Big(\sum_{j \in I - \{i\}} x_{ij}\Big)^2 \ , \ A_3 = \sum_{I^{[2]}} (x_{ij})^2 \ ,$$

$$B_1 = \Big(\sum_{I^{[2]}} y_{ij}\Big)^2 \ , \ B_2 = \sum_{i \in I} \Big(\sum_{j \in I - \{i\}} y_{ij}\Big) \text{ et } B_3 = \sum_{I^{[2]}} (y_{ij})^2 \ .$$

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

$$\begin{aligned} \operatorname{var}(S) = {} & \frac{2}{n^{[2]}} A_3 B_3 + \frac{4}{n^{[3]}} (A_2 - A_3)(B_2 - B_3) \\ & + \frac{1}{n^{[4]}} (A_1 - 4A_2 + 2A_3)(B_1 - 4B_2 + 2B_3) \\ & - \frac{1}{(n^{[2]})^2} A_1 B_1 \ . \end{aligned} \qquad (15)$$

Précisons que le cas symétrique (resp. antisymétrique) pour un codage z est celui où $z_{ij} = z_{ji}$ (resp. $z_{ij} = -z_{ji}$) pour tout (i,j) de $I^{[2]}$. Dans ce dernier cas (symétrique pour x et y ou antisymétrique pour x et y), notre expres-

Si on écrit :

$$\mathrm{var}(S) = \frac{2}{n^{[2]}}\Big(\sum_{I^{[2]}} x_{ij}^2\Big)\Big(\sum_{I^{[2]}} y_{ij}^2\Big)$$

$$+ \frac{4}{n^{[3]}}\Big(\sum_{G} x_{ij}\,x_{ik}\Big)\Big(\sum_{G} y_{ij}\,y_{ik}\Big)$$

$$+ \frac{1}{n^{[4]}}\Big(\sum_{H} x_{ij}\,x_{hk}\Big)\Big(\sum_{H} y_{ij}\,y_{hk}\Big)$$

$$- \Big[\frac{1}{n^{[2]}}\Big(\sum_{I^{[2]}} x_{ij}\Big)\Big(\sum_{I^{[2]}} y_{ij}\Big)\Big]^2 , \qquad (16)$$

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

La correspondance entre notre expression (16) et celle (15') de Mantel est alors claire à partir de l'identification qu'indique les coefficients $\frac{2}{n^{[2]}}$, $\frac{4}{n^{[3]}}$, $\frac{1}{n^{[4]}}$ et $\frac{1}{(n^{[2]})^2}$. Si notre expression [cf. (14) et (16)] est plus claire

d'un point de vue formellement conceptuel, celle de Mantel se prête mieux au calcul puisque les termes qu'elle comprend représentent des sommes doubles.

La forme ainsi réduite de la variance de l'indice brut aléatoire s'applique dans le cas de la comparaison de deux partitions associées à deux variables qualitatives nominales. En effet, cette comparaison peut être assimilée à celle de deux codages symétriques de $O^{[2]}$. En posant — pour une même partition π — $x_{ij} = 1$ (resp. 0) si les objets o_i et o_j sont réunis (resp. séparés) par la partition π, la relation, au lieu d'être considérée au niveau de l'ensemble $O^{[2]}$ des paires, est considérée au niveau de l'ensemble $O^{[2]}$ des couples. L'indice brut et celui aléatoire associé s'en trouvent multipliés par 2 ; la moyenne et la variance s'en trouvant respectivement multipliés par 2 et 4.

La forme réduite de la variance [cf. (15) ou (16)] ne s'applique pas dans le cas de la comparaison de deux préordres totaux associés à deux variables qualitatives ordinales avec le codage que l'on a adopté qui n'est ni symétrique, ni antisymétrique ; ce dernier correspond – pour un même préordre total ω – à poser $x_{ij} = 1$ (resp. 0) si l'objet o_i précède strictement celui o_j (resp. sinon). Toutefois, cette forme réduite de la variance s'applique en considérant le codage antisymétrique d'un même préordre total ω : $x_{ij} = 1$ si o_i précède strictement o_j, $x_{ij} = -1$ si o_i suit strictement o_j et $x_{ij} = 0$ si o_i et o_j se trouvent dans la même classe du préordre.

L'objet de cette étude est de fournir, dans un cadre assez général, une expression de la variance [(15) ou (16)] en vue de déter-

miner la forme limite de l'indice 'centré réduit' (5), dans des conditions asymptotiques qui s'expriment clairement en cas de comparaison de variables qualitatives relationnelles. On aboutira ainsi à une décomposition de la variance, en éléments positifs, chacun d'un ordre fixé par rapport à n. Il s'agira ensuite de « voir » cette expression limite dans chacun des deux cas de comparaison de deux variables qualitatives nominales ou ordinales. On se rendra ainsi compte de ce qui intervient de façon fondamentale dans la normalisation de ces indices. Nous avons déjà mentionné que notre approche permettait d'obtenir comme cas particulier le coefficient $\rho(\alpha,\beta)$ de K. Pearson entre les deux attributs logiques α et β. Pour cette situation, la forme limite de $Q(\alpha,\beta)$ (cf. (5)) est $\sqrt{n}\,\rho(\alpha,\beta)$, où $\rho(\alpha,\beta)$ est un coefficient pur dont la limite ne dépend pas de n. La question se pose de savoir ce qu'il

en est pour la comparaison de deux variables relationnelles dont la représentation est au niveau de $O \times O$, alors que la représentation des attributs logiques conduisant à $Q(\alpha,\beta) = \sqrt{n}\ \rho(\alpha,\beta)$, se fait au niveau de O. Notre analyse formelle nous permettra de répondre à la question.

II - EXPRESSION «FACTORIELLE» DE LA VARIANCE

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

$\{ x(i,j) / (i,j) \in I^{[2]} \}$ et $\{ y(i,j) / (i,j) \in I^{[2]} \}$ les deux codages symétriques (resp. antisymétriques) de $O^{[2]}$, introduisons les paramètres suivants que nous situons par rapport à ceux de Mantel :

$$L_1 = \sum \{ x(i,j) / (i,j) \in I^{[2]} \} = \sqrt{A_1}$$

$$L_2 = \sum \{ [x(i,j)]^2 / (i,j) \in I^{[2]} \} = A_3 \qquad (1)$$

$$U = \sum \{ x(i,j)\, x(i,k) / (i,j,k) \in G \} = (A_2 - A_3)$$

On introduit de même – par rapport au second codage – respectivement, M_1, M_2 et V.

Désignons enfin par

$$\ell_1 = L_1 / n^{[2]} \quad , \ell_2 = L_2 / n^{[2]} \quad , u = U / n^{[3]} \quad , \qquad (2)$$

$$m_1 = M_1 / n^{[2]} \quad , \quad m_2 = M_2 / n^{[2]} \text{ et } v = V / n^{[3]}$$

les paramètres moments factoriels par rapport auxquels nous allons exprimer var (S) en démarrant de l'expression (15) $(\S I)$ qui devient :

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

① ② ③ ④

Ayant numéroté les termes composant la variance, commençons par réduire ③ auquel de proche en proche nous intégrerons les autres termes

$$③ : \frac{1}{n^{[4]}}\left(L_1^2 m_1^2 - 2L_1^2 m_2 - 4L_1^2 V - 2L_2 m_1^2 + 4L_2 m_2 + 8L_2 V - 4U m_1^2 + 8U m_2 + 16 U V\right)$$

On intégrant le terme ④ on a (4)

$$\frac{1}{n^{[2]}} \times \frac{2(2n-3)}{n^{[4]}} L_1^2 m_1^2 + \frac{1}{n^{[4]}} \times \left[4L_2 m_2 + 16UV - 2(L_1^2 m_2 + L_2 m_1^2) + 8(L_2 V + m_2 U) - 4(L_1^2 V + m_1^2 U)\right]$$

Récapitulons une première expression de la variance : (5)

$$\text{var}(S) = 2n^{[2]} l_2 m_2 + 4 n^{[3]} uv + (n^{[2]})^2 \left[\frac{2(2n-3)}{(n-2)(n-3)}\right] l_1^2 m_1^2 + 4\frac{n^{[2]}}{(n-2)(n-3)} l_2 m_2$$

$$+ 16\left[\frac{n(n-1)(n-2)}{(n-3)}\right] uv - 2\left[\frac{(n^{[2]})^2}{(n-2)(n-3)}\right](l_1^2 m_2 + l_2 m_1^2)$$

$$- 4\left[\frac{(n^{[2]})^2}{(n-3)}\right](l_1^2 v + m_1^2 u) + 8\frac{n^{[2]}}{(n-3)}(l_2 v + m_2 u) \qquad (6)$$

Nous voulons rapprocher les deux termes soulignés. Le premier (troisième terme de la somme) s'écrit :

$$4\,\frac{[n(n-1)]^2}{(n-3)}\,\ell_1^2 m_1^2 + 2\,\frac{[n(n-1)]^2}{(n-2)(n-3)}\,\ell_1^2 m_1^2 \qquad (7)$$

Si maintenant, on ne garde de l'expression de la variance que les termes pouvant donner lieu à du $n^{[3]}$, on obtient :

$$4\,n^{[3]}uv + 4\,\frac{[n(n-1)]^2}{(n-3)}\left(\ell_1^2 m_1^2 - \ell_1^2 v - m_1^2 u\right) \qquad (8)$$

qui peut se mettre sous la forme :

$$4\,\frac{[n(n-1)]^2}{(n-3)}\,(u-\ell_1^2)(v-m_1^2) - 8\,\frac{n(n-1)(2n-3)}{(n-3)}\,uv \qquad (9)$$

ou encore :

$$4\,\frac{[n(n-1)]^2}{(n-3)}\,(u-\ell_1^2)(v-m_1^2) - 16\,\frac{[n(n-1)(n-2)]}{(n-3)}\,uv - 8\,\frac{[n(n-1)]}{(n-3)}\,uv \qquad (10)$$

Nous allons maintenant intégrer à cette expression le deuxième terme laissé de (7) ainsi que les autres termes non soulignés de (6). On obtient :

$$\mathrm{var}(S) = 4\frac{[n(n-1)]^2}{(n-3)}(u-\ell_1^2)(v-m_1^2) + \{\,2\frac{[n(n-1)]^2}{(n-2)(n-3)}\ell_1^2 m_1^2 + 2n(n-1)\ell_2 m_2$$

$$- 2\,\frac{[n(n-1)]^2}{(n-2)(n-3)}(\ell_1^2 m_2 + \ell_2 m_1^2)\} - 8\,\frac{[n(n-1)]}{(n-3)}\{uv - (\ell_2 v + m_2 u)\}$$

$$+ \frac{4\,[n(n-1)]}{(n-2)(n-3)}\ell_2 m_2 \,. \qquad (11)$$

où nous avons regroupé les termes selon l'ordre de grandeur.

La deuxième accolade peut se réduire à

$$2\,\frac{[n(n-1)]^2}{(n-2)(n-3)}(\ell_2-\ell_1^2)(m_2-m_1^2) - 4\,\frac{n(n-1)(2n-3)}{(n-2)(n-3)}\ell_2 m_2$$

$$= 2\,\frac{[n(n-1)]^2}{(n-2)(n-3)}(\ell_2-\ell_1^2)(m_2-m_1^2) - 8\,\frac{[n(n-1)]}{(n-3)}\ell_2 m_2 - 4\,\frac{[n(n-1)]}{(n-2)(n-3)}\ell_2 m_2 \,. \qquad (12)$$

En tenant compte de (12), (11) peut s'écrire :

$$\mathrm{var}(S) = 4\frac{[n(n-1)]^2}{(n-3)}(u-\ell_1^2)(v-m_1^2) + 2\,\frac{[n(n-1)]^2}{(n-2)(n-3)}(\ell_2-\ell_1^2)(m_2-m_1^2)$$

$$- 8\,\frac{[n(n-1)]}{(n-3)}(u-\ell_2)(v-m_2) \,. \qquad (13)$$

Finalement,

THEOREME 1 . En tenant compte des notations (1) et (2), var (S) peut s'écrire comme suit :

$$\text{var}(S) = \frac{2[n(n-1)]}{(n-3)}\left\{ 2[n(n-1)](u-\ell_1^2)(v-m_1^2) + \frac{[n(n-1)]}{(n-2)}(\ell_2-\ell_1^2)(m_2-m_1^2) - 4(u-\ell_2)(v-m_2)\right\} \quad . \quad (14)$$

III - FORME LIMITE DE LA VARIANCE ET DU COEFFICIENT D'ASSOCIATION.

La décomposition (14) de la variance n'a pas son terme dominant strictement positif ; en effet, comme on va le voir $(u-\ell_1^2)$ peut tomber en dessous de zéro et $(v-m_1^2)$ positif . Imaginons pour cela que x et y soient deux relations de partition. Nous allons commencer par considérer x comme définissant une partition en h classes de même effectif $k = n/h$ $\left(\frac{k}{n} = \frac{1}{h}\right)$. Dans ces conditions, on a :

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

$$\ell_1 = \ell_2 = h\,k(k-1)/n(n-1) \text{ et } u = h\,k(k-1)(k-2)/n(n-1)(n-2) \text{ . (1)}$$

$$(u - \ell_1^2) = \frac{1}{(n-1)^2(n-2)}\left[(n-1)(k-1)(k-2) - (n-2)(k-1)^2\right] \text{ . (2)}$$

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

$$(u - \ell_1^2) = \frac{1}{(n-1)^2(n-2)}\left[-nk + n + k^2 - k\right]$$

$$= \frac{-n^2}{(n-1)^2(n-2)}\left(1 - \frac{1}{h}\right)\left(\frac{1}{h} - \frac{1}{n}\right) < 0 \text{ . (3)}$$

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 asymptotiques s'expriment plus naturellement par rapport aux moments absolus . Introduisons alors relativement à x (resp. y) la diagonale $\{x_{ii} / 1 \leq i \leq n\}$ (resp. $\{y_{ii} / 1 \leq i \leq n\}$) . On suppose de façon exclusive l'une des deux situations suivantes :

① $x_{ii} = 1$ et $y_{ii} = 1$ pour tout $i = 1, 2, \ldots, n$.

② $x_{ii} = 0$ et $y_{ii} = 0$ pour tout $i = 1, 2, \ldots, n$.

① correspond au cas où les deux relations sont symétriques et ② au cas où les deux relations sont antisymétriques . Dans ces conditions

$$D_1 = \sum_i x_{ii} \quad \text{et} \quad D_2 = \sum_i x_{ii}^2$$ sont égaux (à n ou à 0) entre eux et avec

$$E_1 = \sum_i y_{ii} \quad \text{et} \quad E_2 = \sum_i y_{ii}^2 \quad .$$

D peut désigner la valeur commune (n ou 0) et $d = D/n$ est égal à 1 ou 0. Introduisons

$$\left.\begin{aligned} L_{10} &= L_1 + D = \underbrace{\quad}_{(i,j)} \{x_{ij} \,/\, (i,j) \in I \times I\} \\ L_{20} &= L_2 + D = \underbrace{\quad} \{x_{ij}^2 \,/\, (i,j) \in I \times I\} \\ U_0 &= \underbrace{\quad} \{x_{ij} x_{ik} \,/\, (i,j,k) \in I \times I \times I\} \end{aligned}\right\} \qquad (4)$$

On a

$$U_0 = \underbrace{\quad}_{[i,j,k]} x_{ij} x_{ik} + \underbrace{\quad}_{[i,k]} x_{ii} x_{ik} + \underbrace{\quad}_{[i,j]} x_{ij} x_{ii} + \underbrace{\quad}_{[i,j]} x_{ij}^2 + \underbrace{\quad}_{i} x_{ii}^2,$$

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

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

D'où

$$U = U_0 - 2d L_1 - L_2 - nd = U_0 - 2d(L_{10} - nd) - (L_{20} - nd) - nd$$

$$= U_0 - 2d L_{10} - L_{20} + 2nd^2 \,. \qquad (5)$$

Introduisons à présent les moments absolus :

$$p_1 = L_{10}/n^2 \quad , p_2 = L_{20}/n^2 \quad \text{et} \quad q = U_0/n^3 \quad . \qquad (6)$$

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

$$r_1 = M_{10}/n^2 \quad , r_2 = M_{20}/n^2 \quad \text{et} \quad s = V_0/n^3 \quad . (7)$$

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

Compte tenu de (5), on a

$$u = \frac{n^2}{(n-1)(n-2)} \left[q - \frac{2d}{n} p_1 - \frac{1}{n} p_2 + 2 \frac{d^2}{n^2} \right] , \qquad (8)$$

d'autre part,

$$l_1 = \frac{n}{(n-1)} \left(p_1 - \frac{d}{n} \right) \quad \text{et} \quad l_1^2 = \frac{n^2}{(n-1)^2} \left(p_1^2 - \frac{2}{n} d p_1 + \frac{d^2}{n^2} \right) \quad . (9)$$

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

$$(u - l_1^2) = \frac{n^2}{(n-1)(n-2)} (q - p_1^2) + \frac{1}{(n-2)} (l_1^2 - l_2) \quad . (10)$$

Dans ces conditions,

$$(u-\ell_1^2)(v-m_1^2) = \frac{n^4}{(n-1)^2(n-2)^2}(q-p_1^2)(s-r_1^2)$$

$$+\frac{n^2}{(n-1)(n-2)^2}[(q-p_1^2)(m_1^2-m_2)+(s-r_1^2)(\ell_1^2-\ell_2)]$$

$$+\frac{1}{(n-2)^2}(\ell_1^2-\ell_2)(m_1^2-m_2) \quad . \quad (11)$$

Cette expression est à multiplier par $[4n^2(n-1)^2/(n-3)]$ avant d'intervenir comme le premier terme de la variance. En différenciant les termes par ordre de grandeur, on obtient :

$$\left\{\frac{4n^6}{(n-2)^2(n-3)}(q-p_1^2)(s-r_1^2)+\frac{4n^4(n-1)}{(n-2)^2(n-3)}[(q-p_1^2)(m_1^2-m_2)+(s-r_1^2)(\ell_1^2-\ell_2)]\right.$$

$$\left.+\frac{4n^2(n-1)^2}{(n-2)^2(n-3)}(\ell_1^2-\ell_2)(m_1^2-m_2) \quad . \quad (12)\right.$$

Le deuxième terme de var(S) (cf. (14)) se met sous la forme:

$$\left\{+\frac{2n^2(n-1)^2}{(n-2)(n-3)}(\ell_2-\ell_1^2)(m_2-m_1^2) \quad . \quad (13)\right.$$

Dans le troisième terme de var(S) intervient $(u-\ell_2)(v-m_2)$ que nous allons décomposer comme suit :

$$(u-\ell_2)(v-m_2)=[(u-\ell_1^2)-(\ell_2-\ell_1^2)][(v-m_1^2)-(m_2-m_1^2)]\,. \quad (14)$$

Compte tenu de (10) ci-dessus :

$$(u-\ell_2)=\frac{n^2}{(n-1)(n-2)}(q-r_1^2)-\frac{(n-1)}{(n-2)}(\ell_2-\ell_1^2) \quad . \quad (15)$$

En adoptant une même forme pour $(v-m_2)$, on a

$$\begin{aligned}(u-\ell_2)(v-m_2)=&\frac{n^4}{(n-1)^2(n-2)^2}(q-r_1^2)(s-r_1^2)\\&-\frac{n^2}{(n-2)^2}[(q-r_1^2)(m_2-m_1^2)+(s-r_1^2)(\ell_2-\ell_1^2)]\\&+\frac{(n-1)^2}{(n-2)^2}(\ell_2-\ell_1^2)(m_2-m_1^2) \quad . \quad (16)\end{aligned}$$

Pour obtenir le troisième terme de la variance var(S) [cf. (14) § II], on a à multiplier l'expression (16) par $-8n(n-1)/(n-3)$. On a alors :

$$\begin{cases} -\dfrac{8n^5}{(n-1)(n-2)^2(n-3)}(q-p_1^2)(s-r_1^2) \\ \qquad +\dfrac{8n^3(n-1)}{(n-2)^2(n-3)}\left[(q-p_1^2)(m_2-m_1^2)+(s-r_1^2)(l_2-l_1^2)\right] \\ \qquad -\dfrac{8n(n-1)^3}{(n-2)^2(n-3)}(l_2-l_1^2)(m_2-m_1^2) \end{cases} \quad . (17)$$

Nous avons fait précéder d'une accolade les expressions (12), (13) et (17) qui interviennent dans var (S) par leur somme. Il y a en fait trois expressions composantes de base qui déjà interviennent dans (12) et (13) et qui réaparaissent dans (17). Dans ces conditions, nous allons commencer par reordonner le tout par rapport à ces trois expressions de base qui sont $(q-p_1^2)(s-r_1^2)$, $\left[(q-p_1^2)(m_2-m_1^2)+(s-r_1^2)(l_2-l_1^2)\right]$ et $(l_2-l_1^2)(m_2-m_1^2)$. Ainsi les coefficients ci-dessous concernent le regroupement de (12), (13) et (17).

Le coefficient de $(q-p_1^2)(s-r_1^2)$ est

$$\frac{4n^5}{(n-2)^2(n-3)}\left(n-\frac{2}{n-1}\right)=\frac{4n^5(n+1)}{(n-1)(n-2)(n-3)}\ . \qquad (18)$$

Le coefficient de $[(q-p_1^2)(m_2-m_1^2)+(s-r_1^2)(l_2-l_1^2)]$ est

$$\frac{4n^3(n-1)}{(n-2)^2(n-3)}(-n+2)=-\frac{4n^3(n-1)}{(n-2)(n-3)}\ . \qquad (19)$$

Le coefficient de $(l_2-l_1^2)(m_2-m_1^2)$ est

$$\frac{2n(n-1)^2}{(n-2)(n-3)}\left[\frac{2n}{(n-2)}+n-\frac{4(n-1)}{(n-2)}\right]=\frac{2n(n-1)^2}{(n-3)}\ . \qquad (20)$$

Au facteur $1/(n-3)$ près, considérons la contribution des deux derniers termes :

$$2n(n-1)^2\left\{(l_2-l_1^2)(m_2-m_1^2)-\frac{2n^2}{(n-1)(n-2)}\left[(q-p_1^2)(m_2-m_1^2)+(s-r_1^2)(l_2-l_1^2)\right]\right., \qquad (21)$$

qu'on peut écrire sous la forme :

$$2n(n-1)^2\left[(\ell_2-\ell_1^2)-\frac{2n^2}{(n-1)(n-2)}(q-p_1^2)\right]\left[(m_2-m_1^2)-\frac{2n^2}{(n-1)(n-2)}(s-r_1^2)\right]$$

$$-\frac{8n^5}{(n-2)^2}(q-p_1^2)(s-r_1^2) \quad . \quad (22)$$

Regroupons le dernier terme de (22) avec celui qui correspond à (18). Au facteur $1/(n-3)$, on a comme coefficient multiplicatif :

$$\frac{4n^5}{(n-2)}\left[\frac{(n+1)}{(n-1)}-\frac{2}{(n-2)}\right]=\frac{4n^6(n-3)}{(n-1)(n-2)^2} \quad . \quad (23)$$

On a alors

THEOREME 2

$$\text{var}(S)=\frac{4n^6}{(n-1)(n-2)^2}(q-p_1^2)(s-r_1^2)$$

$$+\frac{2n(n-1)^2}{(n-3)}\left\{\left[(\ell_2-\ell_1^2)-\frac{2n^2}{(n-1)(n-2)}(q-p_1^2)\right]\right.$$

$$\left.\times\left[(m_2-m_1^2)-\frac{2n^2}{(n-1)(n-2)}(s-r_1^2)\right]\right\} \quad . \quad (24)$$

A partir de (24), on établit l'ordre de grandeur de var(S) pour n «grand» :

COROLLAIRE 1 - La forme limite - pour n «grand» - de var(S) est :

$$4\,n^3\left\{(q-h_1^2)(s-r_1^2)+\frac{1}{2n}\left[(h_2-h_1^2)-2(q-h_1^2)\right]\left[(r_2-r_1^2)-2(s-r_1^2)\right)\right. . \quad (25)$$

Nous allons à présent déterminer la forme limite du coefficient $Q(x,y)$ [indice brut centré réduit tel que (5) (§ I)]. Pour cela, considérons le numérateur de ce cœfficient qui se met sous la forme

$$n^{[2]}\,(\gamma - l_1 m_1) \quad , \quad (26)$$

où

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

L'expression limite de (26) comporte comme terme dominant

$$n^2(w-h_1 r_1) \quad , \quad (27)$$

où
$$w=\frac{1}{n^2}\sum\{x_{ij}\,y_{ij}/(i,j)\in I\times I\}.$$

Il en résulte le

THEOREME 3. La forme limite du coefficient $Q(x,y)$ pour p_1, p_2 et q (resp. r_1, r_2 et s) tendant vers des limites finies, est

$$\frac{\sqrt{n}}{2}\times\frac{(w-p_1r_1)}{\sqrt{(q-p_1^2)(s-r_1^2)+\frac{1}{2n}[(p_2-p_1^2)-2(q-p_1^2)][(r_2-r_1^2)-2(s-r_1^2)]}} \tag{28}$$

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 :

$$Q(x,y)\cong\frac{\sqrt{n}}{2}\times\frac{(w-p_1r_1)}{\sqrt{(q-p_1^2)(s-r_1^2)}} \tag{29}$$

On se trouve alors en mesure de répondre à la question posée à la fin

du paragraphe I. Comme pour celui de K. Pearson, le coefficient $Q(x,y)$ entre deux variables relationnelle est également — dans les conditions du théorème 3, pour $(q-p_1^2)(s-r_1^2)\neq 0$, 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 $(q-p_1^2)$ [resp. $(s-r_1^2)$] est nécessairement positif, compte tenu de l'expression (25) de la variance.

IV — COMPARAISON DE DEUX VARIABLES PARTITION.

Pour déterminer la forme limite du coefficient d'association entre deux variables partition π et $\varkappa$, il suffit 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 d'une classe de la partition π (resp. χ) de type $(m_1, m_2, \ldots, m_c, \ldots, m_h)$ [resp. $(n_1, n_2, \ldots, n_k)$]. Désignons par $\pi_c = (m_c/n)$ [resp. $\chi_d = (n_d/n)$] et $\nu_{cd} = (n_{cd}/n)$, où n_{cd} est le cardinal de la classe (c,d) de $\pi \wedge \chi$. On a alors

PROPRIETE 1.

$$\left.\begin{aligned} p_1 = p_2 = \sum_{1 \leqslant c \leqslant h} \pi_c^2 \quad , \quad q = \sum_{1 \leqslant c \leqslant h} \pi_c^3 \\ \left(\text{resp. } r_1 = r_2 = \sum_{1 \leqslant d \leqslant k} \chi_d^2 \quad , \quad s = \sum_{1 \leqslant d \leqslant k} \chi_d^3\right) \\ \text{et} \quad w = \sum \left\{ \nu_{cd}^2 \,/\, 1 \leqslant c \leqslant h \,,\, 1 \leqslant d \leqslant k \right\} \end{aligned}\right\} \qquad (1)$$

En effet, x_{ij} étant une variable booléenne

$$\sum_{(i,j)} x_{ij} = \sum_{(i,j)} x_{ij}^2 = \sum_{1 \leqslant c \leqslant h} m_c^2 \quad , \quad (2)$$

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 $\sum\{x_{ij}x_{ik}/(i,j,k)\}$, d'où la relation :

$$\sum\{x_{ij}x_{ik}/(i,j,k)\} = \sum_{1\leqslant c\leqslant h} m_c^3 \quad , \quad (3)$$

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

$$\sum\{x_{ij}y_{ij}/(i,j)\in I\times I\} = \sum_i \{n_{cd}^2/1\leqslant c\leqslant h, 1\leqslant d\leqslant k\}, \quad (4)$$

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

Un équivalent du théorème 3 du paragraphe précédent peut être énoncé dans ce contexte, avec comme conditions asymptotiques $\{\pi_c/1\leqslant c\leqslant h\}$ (resp. $\{x_d/1\leqslant d\leqslant k\}$) tendant

vers une limite finie. Dans ces conditions, la forme limite de $Q(\pi, x)$ est

$$\frac{\sqrt{n}}{2} x \frac{(w - hr)}{\sqrt{(q-h^2)(s-r^2) + \frac{1}{2n}[(h-h^2) - 2(q-h^2)][(r-r^2) - 2(s-r^2)]}} . \quad (5)$$

Nous allons directement établir la

PROPRIETE 2. $(q - h^2)$ est nul si $\pi_c = 1/h$ pour tout $c = 1, 2, \ldots, h$, autrement $(q - h^2)$ est strictement positif.

Si $\pi_c = 1/h$ pour tout $c = 1, 2, \ldots, h$,

$$h = \sum_{1 \leq c \leq h} \frac{1}{h^2} = \frac{1}{h} \quad \text{et} \quad h^2 = \frac{1}{h^2},$$

$$q = \sum_{1 \leq c \leq h} \frac{1}{h^3} = \frac{1}{h^2} .$$

Prouvons à présent que

$$\sum_{1\leqslant c\leqslant h} \pi_c^3 \geqslant \Big(\sum_{1\leqslant c\leqslant h} \pi_c^2\Big)^2 \qquad (6)$$

Cette inégalité est équivalente à la suivante

$$\Big(\sum_{1\leqslant c\leqslant h} \pi_c^3\Big)\Big(\sum_{1\leqslant c\leqslant h} \pi_c\Big) \geqslant \sum_{c} \pi_c^4 + \sum_{[c,c']} \pi_c^2 \pi_{c'}^2 \quad , \quad (7)$$

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

$$\sum_{[c,c']} \pi_c^3 \pi_{c'} \geqslant \sum_{[c,c']} \pi_c^2 \pi_{c'}^2 \quad . \qquad (8)$$

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

$$\pi_c^3 \pi_{c'} + \pi_{c'}^3 \pi_c \geqslant 2\, \pi_c^2 \pi_{c'}^2 \quad , \qquad (9)$$

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 (9) et donc celle (6) soit stricte.

PROPRIETE 3. $A = [(p - p^2) - 2(q - p^2)]$ est positif.

A se met aussi sous la forme $[(p-q) - (q-p^2)]$. Compte tenu de la décomposition sous jacente à la preuve de la propriété 2

$$(q - p^2) = \underbrace{\sum}_{\{c,c'\}} \pi_c \pi_{c'} (\pi_c - \pi_{c'})^2 \quad , \quad (10)$$

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

$$(p-q) = \underbrace{\sum}_{c} \pi_c^2 (1-\pi_c) = \underbrace{\sum}_{[c,c']} \pi_c^2 \pi_{c'} = \underbrace{\sum}_{\{c,c'\}} \pi_c \pi_{c'} (\pi_c + \pi_{c'}) \quad . \quad (11)$$

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

$$\pi_c + \pi_{c'} = (\pi_c - \pi_{c'}) + 2\pi_{c'} > (\pi_c - \pi_{c'})^2 \qquad (12)$$

et

$$(p-q) > (q-p^2) \, . \qquad (13)$$

On remarquera qu'à l'extrémum, où $\pi_c = \frac{1}{h}$ pour tout c, la valeur de A est $\frac{1}{h}\left(1 - \frac{1}{h}\right)$

Ainsi la décomposition fournie de la variance [sous le signe $\sqrt{\ }$ de (5)] est en éléments positifs.

Toutes choses égales par ailleurs, le coefficient $Q(\pi, \chi)$ [cf. (5)] est d'autant plus grand que la partition π (resp. χ) tend à être en classes de même effectif ; 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 taille. Toutefois deux remarques s'imposent. La première est que l'ordre de grandeur du numérateur – lié à $\{ n_{cd}^2 / 1 \leq c \leq h, 1 \leq d \leq k \}$ [cf. (6) § I] – décroît très sensiblement dès lors que la partition π (resp. χ) tend à être en classes de même effectif. 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 en conséquences de pures constructions.

Ainsi le coefficient $Q(\pi, \chi)/\sqrt{n}$ se trouve marqué par la valeur informative de chacune des deux partitions π et χ. Il s'agit d'une tendance qui peut être sou-

haitable . Si on désire un coefficient ne mettant pas trop en évidence cet effet, on prendra:

$$R(\pi,\chi)=\frac{Q(\pi,\chi)}{\sqrt{Q(\pi,\pi)\,Q(\chi,\chi)}} \qquad (14)$$

V- COMPARAISON DE DEUX VARIABLES PREORDRE TOTAL.

Le coefficient que nous voulons obtenir ici par application de l'expression (28) ne correspond pas à celui détaillé au paragraphe I, mais au codage antisymétrique d'un même préordre total auxquel nous avons fait allusion par la suite. Comme dans le cas de comparaison de deux partitions, il y a lieu de pouvoir reconnaître p_1, p_2, q (resp. r_1, r_2, s) et w. Ici encore, on notera par c (resp. d) l'indice courant d'une classe du préordre total ω (resp. ϖ) de composition $(m_1, m_2, \ldots, m_c, \ldots, m_h)$ [resp. $(n_1, n_2, \ldots, n_d, \ldots, n_k)$]. Désignons par $\omega_c = (m_c/n)$ [resp. $\varpi_c = (n_c/n)$] et

$\nu_{cd} = (n_{cd}/n)$ où n_{cd} est le cardinal de la classe (c,d) de $\omega \wedge \varpi$. On a alors

PROPRIETE 1: $p_1 = 0$, $p_2 = 2 \sum_{c<c'} \omega_c \omega_{c'}$, $q = [2 \sum_{c<c'<c''} \omega_c \omega_{c'} \omega_{c''} + \sum_{c<c'} \omega_c \omega_{c'} (\omega_c + \omega_{c'})]$

(resp. $r_1 = 0$, $r_2 = 2 \sum_{d<d'} \varpi_d \varpi_{d'}$, $s = [2 \sum_{d<d'<d''} \varpi_d \varpi_{d'} \varpi_{d''} + \sum_{d<d'} \varpi_d \varpi_{d'} (\varpi_d + \varpi_{d'})]$).

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

$$I \times I = \sum_{c<c'} I_c \times I_{c'} + \sum_{c'>c} I_{c'} \times I_c + \sum_{c} I_c \times I_c \quad . (2)$$

(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

$$\subset \{ x_{ij}^2 \,/\, (i,j) \in I \times I \} \qquad (3)$$

et on considère la même décomposition (2). 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 expression de la forme

$$\subset \{ x_{ij} x_{ik} \,/\, (i,j,k) \in I \times I \times I \} \qquad (4)$$

et référons nous à la décomposition de $I \times I \times I$ conforme à celle $\{ I_c \,/\, 1 \leqslant c \leqslant h \}$ 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 ses 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 :

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

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

Les seuls ensembles qui contribuent à (4) sont $I_{c'} \times I_c^2$ et $I_c \times I_{c'}^2$. Leur contribution est

$$\sum_{c<c'} m_{c'} m_c (m_c - 1) + \sum_{c<c'} m_c m_{c'} (m_{c'} - 1) \qquad (5)$$

qui contribue à l'expression de q par

$$\sum_{c<c'} \omega_c \omega_{c'} (\omega_c + \omega_{c'}) \quad . \qquad (6)$$

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 $c < c' < c''$.

$$I_c \times I_{c'} \times I_{c''} \longrightarrow m_c\, m_{c'}\, m_{c''}$$

$$I_c \times I_{c''} \times I_{c'} \longrightarrow m_c\, m_{c'}\, m_{c''}$$

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

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

$$I_{c''} \times I_c \times I_{c'} \longrightarrow m_c\, m_{c'}\, m_{c''}$$

$$I_{c''} \times I_{c'} \times I_c \longrightarrow m_c\, m_{c'}\, m_{c''}$$

La contribution globale est donc

$$2 \sum_{c<c'<c''} m_c\, m_{c'}\, m_{c''} \qquad (7)$$

laquelle donne au niveau de q :

$$\sum_{c<c'<c''} \omega_c\, \omega_{c'}\, \omega_{c''} \; . \qquad (8)$$

Il reste maintenant à déterminer w :

Considérons $\mathcal{E}\{x_{ij}\, y_{ij} \,/\, (i,j) \in I \times I\}$ et la décomposition de même type que (2), mais relative à $\mathfrak{S}$:

$$I \times I = \underset{d<d'}{\mathcal{E}} I_d \times I_{d'} + \underset{d>d'}{\mathcal{E}} I_d \times I_{d'} + \underset{d}{\mathcal{E}} I_d \times I_d \quad . \quad (9)$$

Si $(i,j) \in I_c \times I_c$ pour un c, ou $(i,j) \in I_d \times I_d$ pour un d, $x_{ij}\, y_{ij} = 0$.

Etant 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 :

$$(I_c \times I_{c'}) \cap (I_d \times I_{d'}) \longrightarrow n_{cd}\, n_{c'd'}$$

$$(I_c \times I_{c'}) \cap (I_{d'} \times I_d) \longrightarrow -n_{cd'}\, n_{c'd}$$

$$(I_{c'} \times I_c) \cap (I_d \times I_{d'}) \longrightarrow -n_{c'd}\, n_{cd'}$$

$$(I_{c'} \times I_c) \cap (I_{d'} \times I_d) \longrightarrow n_{c'd'}\, n_{cd}$$

La contribution globale est

$$2 n_{cd} n_{c'd'} - 2 n_{cd'} n_{c'd} . \quad (10)$$

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

L'expression (25) du paragraphe III de la variance devient – au facteur $4n^3$ près –

$$qs + \frac{1}{2n} (r_2 - 2q)(r_2 - 2s) \quad (11)$$

où q (resp. s) est essentiellement positif.

PROPRIETE 2 : $B = (p_2 - 2q)$ est positif.

On peut écrire p_2 sous la forme

$$2\sum_{c<c'} \omega_c \omega_{c'} \left(\sum_{c''} \omega_{c''}\right) = 2\sum_{c<c'} \omega_c \omega_{c'} (\omega_c + \omega_{c'}) + 2\sum_{c<c'} \omega_c \omega_{c'} \left(\sum \{\omega_{c''} / c'' \notin \{c, c'\}\}\right) \tag{12}$$

$$= 2\sum_{c<c'} \omega_c \omega_{c'} (\omega_c + \omega_{c'}) + 2\sum_{c''<c<c'} \omega_{c''} \omega_c \omega_{c'} + 2\sum_{c<c''<c'} \omega_c \omega_{c''} \omega_{c'} + 2\sum_{c<c'<c''} \omega_c \omega_{c'} \omega_{c''} . \tag{13}$$

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

$$B = p_2 - 2q = 2\sum_{c<c'<c''} \omega_c \omega_{c'} \omega_{c''} > 0 \tag{14}$$

Ainsi la décomposition (11) est en éléments positifs.

Ici encore on peut confronter le comportement de $Q(\omega,\varpi)$ à celui de

$$R(\omega,\varpi)=\frac{Q(\omega,\varpi)}{\sqrt{Q(\omega,\omega)\,Q(\varpi,\varpi)}} \quad . \quad (15)$$

REFERENCES.

[1] Daniels, H.E. (1944), 'The relation between measures of correlation in the universe of sample permutations', Biometrika, vol. 33, 129-135.

[2] Lecalvé, G. (1976), 'Un indice de similarité pour des variables de types quelconques', Stat. et Anal. des Données, 01-02, 39-47.

[3] Lerman, I.C. (1976), 'Formal analysis of a general notion of proximity between variables', Congrès Européen des Statisticiens, Grenoble, published by North Holland in 1977.

[4] Lerman, I.C. (1973), 'Etude distributionnelle de statistiques de proximité entre structures finies de même type ; application à la classification automatique', Cahiers du B.U.R.O. n°19, Paris.

[5] Lerman, I.C. (1981), 'Classification et analyse ordinale des données', Dunod, Paris.

[6] Lerman, I.C. (1983_a), 'Indices d'association partielle entre variables qualitatives nominales', R.A.I.R.O., série R.O. vol 17, n°3, 213-259.

[7] Lerman, I.C. (1983_b), 'Indices d'association partielle entre variables qualitatives ordinales', Publ. I.S.U.P., XXVIII, fasc 1,2,7-46.