Interpretation non lineaire d'un coefficient d'association entre modalites d'une juxtaposition de tables de contingence

Israël-César Lerman

IRIA

CENTRE DE RENNES

IRISA

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél. 954 90 20

Rapports de Recherche

N° 179

INTERPRÉTATION NON LINÉAIRE D'UN COEFFICIENT D'ASSOCIATION ENTRE MODALITÉS D'UNE JUXTAPOSITION DE TABLES DE CONTINGENCE

Israël César LERMAN

Décembre 1982

INTERPRETATION NON LINEAIRE D'UN COEFFICIENT
D'ASSOCIATION ENTRE MODALITES D'UNE JUXTAPOSITION
DE TABLES DE CONTINGENCE

Israël César LERMAN

Publication Interne n°182 - Novembre 82 - 34 pages

RESUME : La structure du tableau des données est celle d'une juxtaposition "horizontale" de tables de contingence indexée par $I\times(J^{(1)}\cup\ldots\cup J^{(\ell)}\cup\ldots\cup J(L))$ où chaque $I\times J^{(\ell)}, 1\leq\ell\leq L$, est une table de contingence. Le problème est celui de la définition d'un indice d'association entre éléments de $J = \cup\{J^{(\ell)}/1\leq\ell\leq L\}$ dont l'élaboration doit être justifiée que les modalités à comparer soient ou non exclusives. Cette justification est obtenue dans le cadre d'une approche classiquement utilisée par l'auteur (représentation ensembliste des variables -indice brut adéquat de proximité-centrage et réduction de cet indice relativement à une hypothèse d'absence de lien qui respecte les structures cardinales des variables à comparer). L'indice obtenu est comparé avec un indice précédemment proposé dont la forme corrélative repose sur la représentation géométrique de I à travers J telle qu'elle est fournie dans l'analyse des correspondances. Les deux indices conduisent au même coefficient de K. Pearson dans le cas d'un tableau "disjonctif-complet". Enfin, on étudie la condition à laquelle l'indice obtenu permet la construction sans inversions d'un arbre binaire de classification sur J au moyen de l'algorithme classique de classification ascendante hiérarchique.

SUMMARY : We are concerned here by a structure of data which may be represented by an "horizontal" juxtaposition of contingency tables : $I\times(J^{(1)}\cup\ldots\cup J^{(\ell)}\cup\ldots\cup J^{(L)})$ where each $I\times J^{(\ell)}, 1\leq\ell\leq L$, represents a contingency table. One goal of this paper is to built a proximity index between two elements j and j' of $J=\cup\{J^{(\ell)}/1\leq\ell\leq L\}$ which must be justified in both cases where j and j' are exclusive or not. This justification is obtained in the context of the approach classicaly used by the author (set representation of the descriptive variables -suitable rough index of proximity- standardisation of this index with respect to hypothesis of non link which respects the cardinality structures of the variables to be associated). We compare the resulting index with a preceeding one obtained for the same structure of the data and having correlative form, but based on geometrical representation of I through J. Both indices lead to the K. Pearson's coefficient of association in the case where I represents the set of the individuals or objects ("disjonctive complete incidence table"). Finally we analyse a necessary condition for which the new index obtained permits the construction without inversions of a binary classification tree on J by a way of a classical algorithm.

INTERPRETATION NON LINEAIRE D'UN COEFFICIENT D'ASSOCIATION ENTRE MODALITES D'UNE JUXTAPOSITION DE TABLES DE CONTINGENCE.

I- INTRODUCTION.

La donnée à laquelle nous nous intéresserons plus particulièrement ici est définie par une juxtaposition « horizontale » de tables de contingences, indéxée par un ensemble de la forme

$$I \times (J^{(1)} \cup J^{(2)} \cup \cdots \cup J^{(\ell)} \cup \cdots \cup J^{(L)}), \qquad (1)$$

où I (resp. $J^{(\ell)}$, $1 \leqslant \ell \leqslant L$) se trouve défini par l'ensemble des modalités d'une variable-partition ; en d'autres termes, I et chaque $J^{(\ell)}$ est un système exhaustif de modalités <u>exclusives</u>.

On considère le problème de la comparaison, <u>relativement à I</u>, des éléments deux à deux de l'ensemble suivant J des modalités :

$$J = J^{(1)} \cup J^{(2)} \cup \cdots \cup J^{(\ell)} \cup \cdots \cup J^{(L)}. \qquad (2)$$

Alors que deux modalités j_ℓ et j'_ℓ d'un même $J^{(\ell)}$ sont exclusives, il n'en est généralement pas de même de deux modalités j_ℓ et $j_{\ell'}$ appartenant respectivement à deux ensembles distincts $J^{(\ell)}$ et $J^{(\ell')}$ $(\ell \neq \ell')$. Un aspect de notre préoccupation qui a conduit à cet article consiste précisément à justifier l'usage d'un même coefficient d'association aussi bien pour comparer j_ℓ et j'_ℓ que pour comparer j_ℓ et $j_{\ell'}$ et ce, à travers I.

Il est inutile d'insister sur l'importance pratique de la structure des données qui nous concerne ici et qu'on rencontre

fréquemment, notamment en Géographie Sociale, Economie, Linguistique, ... D'autre part, lorsque la partition indéxant les lignes se réduit à la partition discrète, I représente l'ensemble des individus ou objets et le tableau des données est dans ce cas communément appelé "tableau disjonctif complet" ; on le rencontre comme résultat d'un questionnaire.

En commençant par considérer une seule table de contingence I x J, I. G. Lerman et B. Tallur [TALLUR (1978), LERMAN et TALLUR (1980)] ont défini un indice d'association entre éléments de J ayant une forme corrélative.

Cette définition prenait d'une part en compte la représentation euclidienne de I à travers J telle qu'elle est fournie dans l'analyse des correspondances et d'autre part, le fait que l'application de notre démarche de construction d'un indice d'association entre variables, conduit lorsque ces dernières sont quantitatives, à un facteur multiplicatif constant près, au coefficient de corrélation entre les deux variables.

Dans cette démarche que nous rappelerons de façon plus précise ci-dessous, on adopte une représentation ensembliste des variables et on introduit une hypothèse d'absence de lien par rapport à laquelle on se réfère. L'objet principal de ce papier est de montrer que cette seule démarche de construction permet, sans aucune référence à une représentation euclidienne, de retrouver le même type d'indice et de le généraliser très naturellement au cas où J est de la forme (2) ci-dessus. Cette généralisation nous permettra de «comprendre», pour ce qui concerne la comparaison entre deux modalités non exclusives, la différence entre un coefficient d'association <u>totale</u> et d'association <u>relative</u> à l'ensemble I des modalités d'une variable qualitative nominale.

II. LES DEUX REPRESENTATIONS DE LA CORRELATION.

Pour la comparaison de deux variables d'un même type,

nous allons considérer les deux cas les plus simples ; le premier est celui où les deux variables sont quantitatives numériques et le second est celui où il s'agit d'attributs de description.

1 - Cas où les deux variables sont numériques.

Désignons par (v, w) les deux variables et par $\{x_i / 1 \leq i \leq n\}$ (resp. $\{y_i / 1 \leq i \leq n\}$) la suite des valeurs de la variable v (resp. w) sur la suite des individus. Le coefficient de corrélation entre les deux variables

$$\rho(v,w) = \frac{\text{moy}\left[(v - \text{moy}(v))(w - \text{moy}(w))\right]}{\sqrt{\text{var}(v)\,\text{var}(w)}}, \qquad (1)$$

où moy. (resp. var.) désigne la moyenne (resp. variance) sur l'ensemble que nous noterons I des individus, est interprété selon le point de vue géométrique en analyse des données, comme le cosinus de l'angle de deux vecteurs.

De façon précise si X (resp. Y) désigne le vecteur de $\mathbb{R}^n$ dont la suite des composantes est $(x_i / 1 \leq i \leq n)$ (resp. $(y_i / 1 \leq i \leq n)$) et si $\mathbb{1}$ désigne le vecteur de $\mathbb{R}^n$ dont toutes les composantes sont égales à 1, $\rho(v,w)$ est le cosinus de l'angle des deux vecteurs suivants :

$$\left.\begin{aligned} X' &= X - \frac{1}{n}\langle X, \mathbb{1}\rangle\, \mathbb{1} \\ Y' &= Y - \frac{1}{n}\langle Y, \mathbb{1}\rangle\, \mathbb{1} \end{aligned}\right\} \qquad (2),$$

où $\langle X, \mathbb{1}\rangle$ désigne le produit scalaire ordinaire, résultant des projections orthogonales des vecteurs X et Y sur l'hyperplan perpendiculaire au vecteur dont toutes les

composantes sont égales à 1.

Une des origines de notre démarche dans l'évaluation des proximités entre variables discrètes et, plus généralement, entre structures statistiques de même type, est que l'indice de corrélation $\rho(v, w)$ peut, au coefficient $1/\sqrt{n-1}$ près, être obtenu de la manière suivante.

On commence par introduire un indice "brut" d'association entre les deux variables v et w :

$$s(v, w) = \sum_{1 \leq i \leq n} v(i) w(i) \qquad (3)$$

On considère ensuite une "hypothèse d'absence de lien" (h.a.l.) où on introduit une permutation aléatoire σ dans l'ensemble G_n, muni d'une probabilité uniformément répartie, des $n!$ permutations sur $(1, 2, \ldots, i, \ldots, n)$.

L'h.a.l. peut dans ce cas avoir une forme unilatérale en fixant v (resp. w) et en associant à w (resp. v) une variable aléatoire (v.a.) w' (resp. v') pour laquelle

$$w'(i) = w[\sigma(i)] \quad \text{pour tout } i = 1, 2, \ldots, n$$

$$\left(\text{resp.} \quad v'(i) = v[\sigma(i)] \quad \text{pour tout } i = 1, 2, \ldots, n\right).$$

A $s(v, w)$ se trouvent associées les deux v.a. duales et de même loi asymptotiquement normale [WALD & WOLFOWITZ (1944), NOEUTHER (1949), HÁJEK (1961)], suivantes :

$$\left.\begin{aligned} S(v, w') &= \sum_{1 \leq i \leq n} v(i)\, w[\sigma(i)] \\ \text{et} \quad S(v', w) &= \sum_{1 \leq i \leq n} v[\sigma(i)]\, w(i) \end{aligned}\right\} \qquad (4)$$

La moyenne et la variance communes sont respectivement égales à

$$n \, moy(v) \, moy(w) \quad et \quad \frac{n^2}{(n-1)} var(v) \, var(w) \quad ; \quad (5)$$

de sorte que

$$\frac{s(v,w) - \mathcal{E}(S)}{var(S)} = \sqrt{n-1} \; \rho(v,w) \; . \quad (6)$$

où S est l'une des deux v.a. (4).

Ce résultat élémentaire est ce que j'appelerai une interprétation "non linéaire" du coefficient de corrélation entre variables numériques.

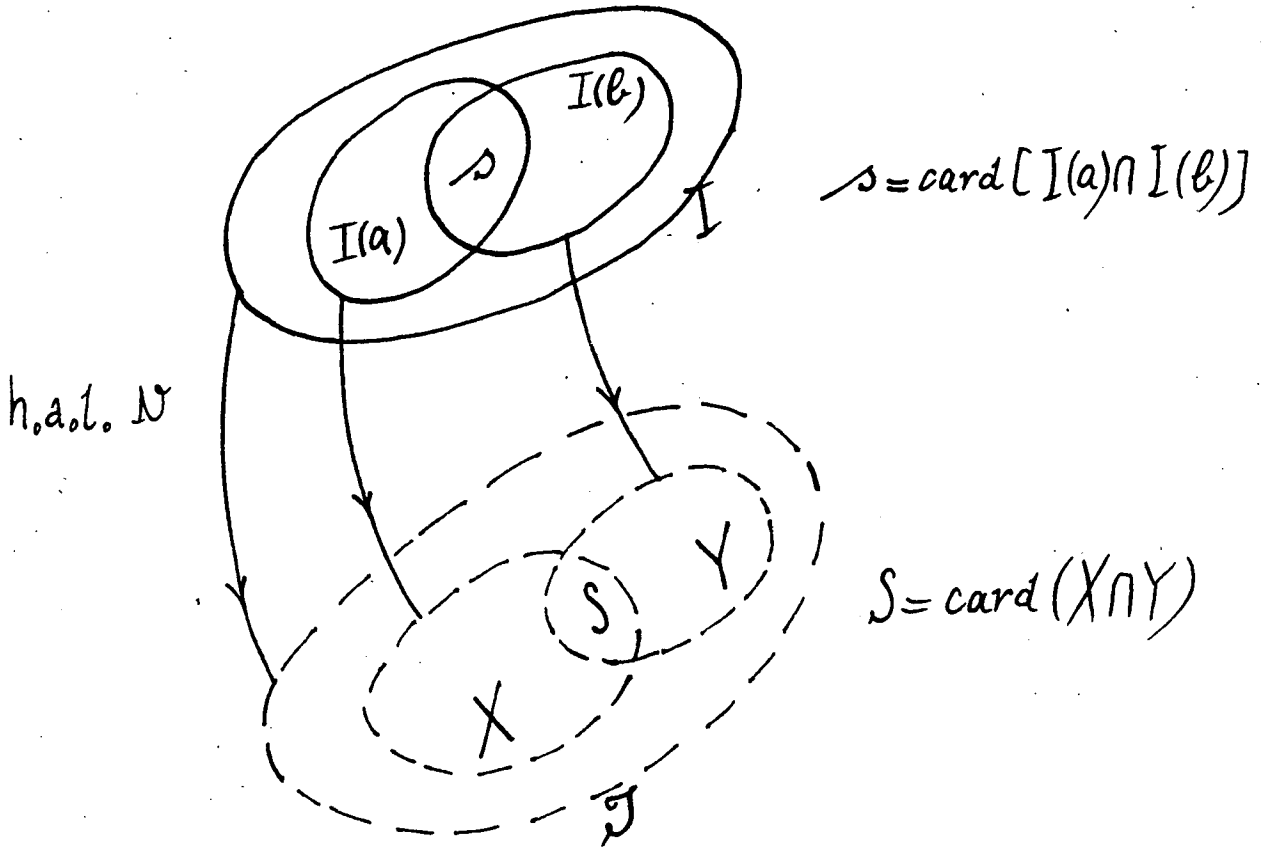
2- Cas où les deux variables sont des attributs.

Désignons par (a, b) le couple d'attributs descriptifs à comparer, par $(I(a), I(b))$ le couple de parties de I, où $I(a)$ (resp. $I(b)$) est l'ensemble des individus qui possèdent l'attribut a (resp. b) et par $(\varepsilon(a), \varepsilon(b))$ le couple de vecteurs logiques indiquant respectivement $I(a)$ et $I(b)$: $\varepsilon(a) = (\alpha_1, \alpha_2, \ldots, \alpha_i, \ldots, \alpha_n)$ et $\varepsilon(b) = (\beta_1, \beta_2, \ldots, \beta_i, \ldots, \beta_n)$, où α_i (resp. β_i) est égal à 0 ou 1 selon que l'attribut a (resp. b) est absent ou présent chez l'individu i, $i \in I$.

Le point de vue géométrique conduit à représenter l'attribut a (resp. b) par le point de $\mathbb{R}^n$ dont la suite des composantes est définie par α_i (resp. β_i), $1 \leq i \leq n$, et à opérer comme dans le cas où les variables sont numériques.

Relativement à notre point de vue de représentation ensembliste des variables, les attributs a et b sont figurés par deux points de l'ensemble $\mathcal{P}(I)$ des

parties de I. La situation peut être naïvement shématisée comme suit



Après avoir introduit l'indice brut $s = \mathrm{card}[I(a) \cap I(b)]$, on considère une h.a.l. N qui associe au triplet $\{I ; I(a), I(b) / I(a) \subset I, I(b) \subset I\}$, un triplet d'ensembles aléatoires $\{J ; X, Y / X \subset J, Y \subset J\}$.

L'h.a.l. doit d'une « certaine façon » respecter les caractéristiques cardinales de $I(a)$, $I(b)$ et I et à cet égard, nous avons pu dégager trois formes fondamentales de l'h.a.l. N : N_1, N_2 et N_3 [LERMAN (1981a) chap. 2].

Pour N_1, $J = I$ et X (resp. Y) est un élément aléatoire dans l'ensemble $\mathcal{P}_{n(a)}(I)$ (resp. $\mathcal{P}_{n(b)}(I)$), muni d'une probabilité uniformément répartie, des parties de I de même cardinal $n(a) = \mathrm{card}(I(a))$ (resp. $n(b) = \mathrm{card}(I(b))$).

D'autre part, X et Y sont indépendants. Dans ces conditions la v.a. $S = \mathrm{card}(X \cap Y)$ est hypergéométrique de moyenne $n(a)n(b)/n$ et de variance $n(a)n(\bar{a})n(b)n(\bar{b})/n^2(n-1)$. L'indice centré réduit est, au coefficient $\sqrt{n-1}$ près, l'indice d'association de K. Pearson.

Pour N_2, $\mathcal{J} = I$. Le choix de X (resp. Y) se fait selon un modèle aléatoire à deux pas :

— le premier consiste dans le choix d'un niveau k (resp. h) de $\mathcal{P}(I)$ avec une probabilité binomiale

$$\binom{n}{k} \alpha^k (1-\alpha)^{n-k}, \text{ où } \alpha = n(a)/n \text{ et } 0 \leq k \leq n,$$

(resp. $\binom{n}{h} \beta^h (1-\beta)^{n-h}$, où $\beta = n(b)/n$ et $0 \leq h \leq n$.)

— le deuxième pas consiste dans le choix uniformément au hasard d'un élément qui est un k-sous ensemble (resp. h-sous ensemble) de I, à ce niveau.

De plus X et Y sont indépendants. Dans ces conditions, on démontre que $S = \mathrm{card}(X \cap Y)$ est une v.a. binomiale de paramètre $\pi = \alpha\beta$. Sa moyenne est donc la même que dans le cas de l'h.a.l. N_1, mais sa variance est égale à $n\pi(1-\pi)$.

Pour N_3, le choix de X (resp. Y) se fait selon un modèle aléatoire à trois pas :

— le premier consiste à associer à I un ensemble aléatoire $\mathcal{J}$, mais où l'aléa ne concerne que la cardinalité de $\mathcal{J}$. On suppose que $\nu = \mathrm{card}(\mathcal{J})$ est une v.a. de Poisson de paramètre $n = \mathrm{card}(I)$:

$$\Pr\{\nu = \ell\} = \frac{n^\ell}{\ell!} e^{-n}$$

— pour $\mathcal{U} = E_0$ fixé, les deux autres pas sont analogues

à ceux de N_2. Les parties aléatoires X et Y étant indépendantes, on démontre que la v.a. $S = \mathrm{card}(X \cap Y)$ suit une loi de Poisson de paramètre $n(a)\,n(b)/n$. On voit que la moyenne de S est la même que dans les deux cas précédents ; mais la variance devient ici égale à $n(a)\,n(b)/n$.

Le deuxième point de vue de représentation ensembliste des variables de description se généralise de façon naturelle pour la comparaison de deux variables qualitatives de toutes sortes [LERMAN (1981 a) Chap. 2]. Nous allons montrer qu'il suffit pour la comparaison de lignes ou colonnes d'une juxtaposition « horizontale » de tableaux de contingence (cf. (1) § I).

III. CAS D'UNE JUXTAPOSITION DE TABLES DE CONTINGENCE.

III.1. CAS D'UN SEUL TABLEAU DE CONTINGENCE.

1. Représentation géométrique.

Soit le tableau de contingence indéxé par $I \times J$:

$$\{k_{ij} \,/\, (i,j) \in I \times J\} \qquad (1)$$

où k_{ij} désigne le nombre d'individus possédant les modalités i de I et j de J. On rappele les notations :

$$\left.\begin{array}{l} k_{i.} = \sum \{k_{ij} \,/\, j \in J\}, \quad k_{.j} = \sum \{k_{ij} \,/\, i \in I\} \\ k_{..} = \sum \{k_{i.} \,/\, i \in I\} = \sum \{k_{.j} \,/\, j \in J\} \end{array}\right\} \quad (2)$$

Nous allons commencer par rappeler le principe de la comparaison de deux modalités j et j' de J, qui repose sur la représentation euclidienne de I à travers J. Cette représen-

tation fournie dans le cadre de l'analyse des correspondances où on associe à chaque i de I, le point de $\mathbb{R}^{|J|}$ dont la suite des coordonnées est $\{k_{ij}/k_{i\cdot} \mid j \in J\}$; il s'agit du « profil » de i à travers J.

Dès lors, chaque j de J se trouve assimilé à une variable quantitative – puisqu'elle est représentée par une forme linéaire coordonnée – dont la valeur sur le i-ème élément est égale à $(k_{ij}/k_{i\cdot})$. D'autre part, l'élément i, pour tout i de I, est affecté du poids $k_{i\cdot}/k_{\cdot\cdot}$.

Dans ces conditions, le coefficient de corrélation entre j et j' de J, se met sous la forme

$$\rho(j,j') = \frac{\sum \left\{ \frac{k_{i\cdot}}{k_{\cdot\cdot}} \left(\frac{k_{ij}}{k_{i\cdot}} - \frac{k_{\cdot j}}{k_{\cdot\cdot}} \right) \left(\frac{k_{ij'}}{k_{i\cdot}} - \frac{k_{\cdot j'}}{k_{\cdot\cdot}} \right) \Big/ i \in I \right\}}{\sqrt{\sum_i \frac{k_{i\cdot}}{k_{\cdot\cdot}} \left(\frac{k_{ij}}{k_{i\cdot}} - \frac{k_{\cdot j}}{k_{\cdot\cdot}} \right)^2 \sum_i \frac{k_{i\cdot}}{k_{\cdot\cdot}} \left(\frac{k_{ij'}}{k_{i\cdot}} - \frac{k_{\cdot j'}}{k_{\cdot\cdot}} \right)^2}} \qquad (3)$$

ou encore, en introduisant les proportions

$(\forall (i,j) \in I \times J)$, $p_{i\cdot} = k_{i\cdot}/k_{\cdot\cdot}$, $p_{\cdot j} = k_{\cdot j}/k_{\cdot\cdot}$ et

$$f_j^i = k_{ij}/k_{i\cdot} ,$$

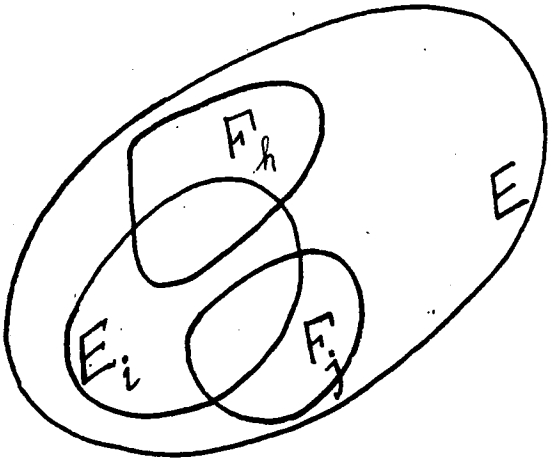
$$\rho(j,j') = \frac{\sum \left\{ p_{i\cdot} (f_j^i - p_{\cdot j})(f_{j'}^i - p_{\cdot j'}) \Big/ i \in I \right\}}{\sqrt{\sum_i p_{i\cdot} (f_j^i - p_{\cdot j})^2 \sum_i p_{i\cdot} (f_{j'}^i - p_{\cdot j'})^2}} \qquad (3')$$

2 – Représentation ensembliste.

Désignons par E l'ensemble des individus ou objets et par $\{E_i / i \in I\}$ (resp. $\{F_j / j \in J\}$) la partition sur E

définie par la variable qualitative nominale dont les modalités indéxent les lignes (resp. colonnes) du tableau de contingence.

Si j et h sont les deux modalités exclusives de J à comparer, relativement à I, la situation peut être, relativement à un même i, schématisée comme suit :



2.1. Indice brut d'association.

Relativement à la classe E_i, le lien brut entre les deux parties disjointes F_j et F_h sera mesuré par

$$\left.\begin{aligned} \varepsilon(j,h/i) &= \frac{n(i\wedge j)\,n(i\wedge h)}{n(i)} \\ &= n(i)\,\frac{n(i\wedge j)}{n(i)} \times \frac{n(i\wedge h)}{n(i)} \end{aligned}\right\} \quad (4)$$

Dans ces conditions, le lien brut entre F_j et F_h relativement à la partition $\{E_i / i \in I\}$, se trouve défini par

$$\mathbb{E}\left\{\frac{n(i\wedge j)\,n(i\wedge h)}{n(i)} \Big/ i\in I\right\} = \mathbb{E}\left\{n(i)\,\frac{n(i\wedge j)}{n(i)} \times \frac{n(i\wedge h)}{n(i)} \Big/ i\in I\right\}, \quad (5)$$

où nous notons ici $n(i) = \mathrm{card}(E_i)$, $n(i\wedge j) = k_{ij} =$

$\mathrm{card}(E_i \cap F_j)$ (resp. $n(i \wedge h) = k_{ih} = \mathrm{card}(E_i \cap F_h)$).

2.2 – L'hypothèse d'absence de lien.

Pour un même i l'h.a.l. entre j et h sera relative à la classe E_i ; son rôle est de détruire le lien entre F_j ou (non exclusif) F_h et E_i, $1 \leq i \leq |I|$.

Une façon naturelle de procéder, conformément à une h.a.l. de même nature que N_1 (cf § II.2), consiste à fixer F_j et F_h et à associer à la partition $\{E_i \,/\, 1 \leq i \leq |I|\}$, une partition aléatoire $\{X_i \,/\, 1 \leq i \leq |I|\}$ dans l'ensemble $\mathcal{P}(n\,;t)$, muni d'une probabilité uniforme, des partitions en classes étiquetées de même type $t = [n(1), n(2), \ldots, n(i), \ldots, n(I)]$.

Nous désignerons par ε_i, ξ_i, φ_j et φ_h les fonctions indicatrices respectives de E_i, X_i, F_j et F_h, $1 \leq i \leq |I|$.

La v.a. associée à l'indice brut (5) peut dans ces conditions se mettre sous la forme

$$E\left\{\frac{1}{n(i)}\left[\sum_{x \in E} \xi_i(x)\,\varphi_j(x)\right]\left[\sum_{y \in E} \xi_i(y)\,\varphi_h(y)\right] \Big/ 1 \leq i \leq |I|\right\}$$

$$= E\left\{\frac{1}{n(i)}\left(E\left\{\xi_i(x)\,\xi_i(y)\,\varphi_j(x)\,\varphi_h(y) \Big/ (x,y) \in E \times E\right\}\right) \Big/ 1 \leq i \leq |I|\right\}. \quad (6)$$

L'indice brut (5) et la v.a. (6) peuvent respectivement être notés $s(j,h\,/\,P)$ et $S(j,h\,/\,P)$, où P désigne la partition $\{E_i \,/\, i \in I\}$.

Nous allons calculer l'espérance mathématique et la variance de la v.a. $S(j,h\,/\,P)$.

2.3 – Moyenne de la v.a. $S(j,h\,/\,P)$.

On peut remarquer que la somme la plus interne de (6) se

réduit à l'expression suivante

$$\sum \{ \xi_i(x)\, \xi_i(y)\, \varphi_j(x)\, \varphi_h(y) \,/\, (x,y) \in F_j \times F_h \} \quad . \quad (7)$$

De plus, pour $x \neq y$, on a

$$\mathcal{E}\left(\xi_i(x)\, \xi_i(y) \right) = \frac{n(i)[n(i)-1]}{n(n-1)} \quad ; \quad (8)$$

il s'agit en effet de la proportion $\binom{n-2}{n(i)-2} \Big/ \binom{n}{n(i)}$ de parties X_i qui contiennent la paire d'objets $\{x,y\}$.

Dans ces conditions, on a

$$\mathcal{E}[S(j,h/P)] = \sum \left\{ \frac{1}{n(i)} \times \frac{n(i)[n(i)-1]}{n(n-1)} \times n(j)\, n(h) \,/\, 1 \leq i \leq |I| \right\}$$

$$= \frac{(n-|I|)}{(n-1)} \times \frac{n(j)\, n(h)}{n} \quad . \quad (9)$$

Il ne faut pas s'étonner de constater que cette moyenne est nulle si la partition P est "discrète" ; en effet, dans ce cas, l'indice brut qui se réduit à $n(j \wedge h)$ est lui même nul puisque j et h sont deux modalités exclusives. Généralement, pour un tableau de contingence courant, $|I|$ est très petit devant n ; de sorte que l'expression (9) peut être approchée par $n(j)\, n(h)/n$. En admettant cette approximation, on peut se rendre compte que le numérateur du coefficient (3) ci-dessus est, au facteur $1/n$ près, l'indice centré

$$\sum_{1 \leq i \leq |I|} \frac{n(i \wedge j)\, n(i \wedge h)}{n(i)} - \frac{n(j)\, n(h)}{n} \qquad (10)$$

Nous allons à présent procéder au calcul de la variance de la v.a. $S(j,h/P)$.

2.4 – Variance de la v.a. $S(j,h/P)$.

Nous allons commencer par calculer le moment absolu d'ordre 2 de la v.a. (6) dont le carré se met sous la forme

$$\sum_{1\leq i\leq |I|}\left\{\frac{1}{n(i)^2}\left(\sum\left\{\xi_i(x)\xi_i(y)\,/\,(x,y)\in F_j\times F_h\right\}\right)^2\right\}$$

$$+2\sum_{1\leq i<i'\leq |I|}\left\{\frac{1}{n(i)n(i')}\left(\sum\left\{\xi_i(x)\xi_i(y)\,/\,(x,y)\in F_j\times F_h\right\}\right)\right.$$

$$\left.\times\left(\sum\left\{\xi_{i'}(x')\xi_{i'}(y')\,/\,(x',y')\in F_j\times F_h\right\}\right)\right\}. \tag{11}$$

La structure d'un couple de couples d'objets $((x,y),(x',y'))$ se trouve définie relativement à la répétition de composantes du premier couple dans le second couple, dans la même position ou non. Des lettres différentes indiquant des objets distincts, cette structure revêt, dans le cas où les deux composantes d'un même couple sont distincts, les sept formes suivantes :

$$\begin{aligned}
&\big((x,y),(x,y)\big)\ ,\ \big((x,y),(y,x)\big),\\
&\big((x,y),(x,t)\big)\ ,\ \big((x,y),(z,x)\big),\\
&\big((x,y),(z,y)\big)\ ,\ \big((x,y),(y,t)\big),\\
&\big((x,y),(z,t)\big).
\end{aligned} \tag{12}$$

Comme d'usage [LERMAN (1981$_a$ Chap. 2) et (1981$_b$)], nous désignerons par

Δ l'ensemble des couples de couples de la forme $\big((x,y),(x,y)\big)$,

$\bar{\Delta}$ " " " " " " " " $\big((x,y),(y,x)\big)$,

G_1 " " " " " " " " $\big((x,y),(x,t)\big)$,

$\bar{G}_1$ " " " " " " " " $\big((x,y),(z,x)\big)$,

G_2 l'ensemble des couples de couples de la forme $((x,y),(z,y))$,
G'_2 " " " " " " " " $((x,y),(y,t))$ et
H " " " " " " " " $((x,y),(z,t))$.

Le calcul (11) va devoir se décomposer conformément à la partition $\{\Delta, \bar{\Delta}, G_1, G'_1, G_2, G'_2, H\}$ de $E^{[2]} \times E^{[2]}$ où $E^{[2]}$ est l'ensemble des couples d'objets distincts. Nous désignerons par

$D = \Delta \cap [(F_j \times F_h) \times (F_j \times F_h)]$ de cardinal $n(j)\,n(h)$,

$B_1 = G_1 \cap [(F_j \times F_h) \times (F_j \times F_h)]$ " " $n(j)\,n(h)[n(h)-1]$,

$B_2 = G_2 \cap [(F_j \times F_h) \times (F_j \times F_h)]$ " " $n(j)[n(j)-1]\,n(h)$ et

$C = H \cap [(F_j \times F_h) \times (F_j \times F_h)]$ " " $n(j)[n(j)-1]\,n(h)[n(h)-1]$.

Chacune des intersections de $(F_j \times F_h) \times (F_j \times F_h)$ avec $\bar{\Delta}$, G'_1 et G'_2 est vide ; d'ailleurs, on peut vérifier que la somme des cardinaux des ensembles D, B_1, B_2 et C, est bien égale à $[n(j)\,n(h)]^2$.

2.4.1. Calcul de l'espérance mathématique de

$$[E\{\xi_i(x)\,\xi_i(y) \,/\, (x,y) \in F_j \times F_h\}]^2.$$

Nous allons commencer par décomposer, conformément à la partition qu'on vient de définir de $(F_j \times F_h) \times (F_j \times F_h)$, le calcul de l'expression en titre de ce sous paragraphe. Cette dernière prend la forme suivante :

$$\mathcal{E}\{\xi_i(x)\xi_i(y) / (x,y) \in F_j \times F_h\}$$
$$+ \mathcal{E}\{\xi_i(x)\xi_i(y)\xi_i(t) / ((x,y),(x,t)) \in B_1\}$$
$$+ \mathcal{E}\{\xi_i(x)\xi_i(y)\xi_i(z) / ((x,y),(z,y)) \in B_2\}$$
$$+ \mathcal{E}\{\xi_i(x)\xi_i(y)\xi_i(z)\xi_i(t) / ((x,y),(z,t)) \in C\}. \qquad (13)$$

Nous allons calculer l'espérance mathématique du terme courant de chacune des sommes précédentes.

α - $\underline{\mathcal{E}\{\xi_i(x)\xi_i(y) / (x,y) \in F_j \times F_h\}}$.

Il s'agit de la proportion de parties X_i de cardinal $n(i)$, qui contiennent les deux objets x et y. Cette proportion est égale à

$$\frac{n(i)[n(i)-1]}{n(n-1)} . \qquad (14)$$

β_1 - $\underline{\mathcal{E}\{\xi_i(x)\xi_i(y)\xi_i(t) / ((x,y),(x,t)) \in B_1\}}$.

Il s'agit de la proportion de parties X_i de cardinal $n(i)$, qui contiennent les trois objets x, y et t. Cette proportion est égale à

$$\frac{n(i)[n(i)-1][n(i)-2]}{n(n-1)(n-2)} . \qquad (15)$$

β_2 - $\underline{\mathcal{E}\{\xi_i(x)\xi_i(y)\xi_i(z) / ((x,y),(z,y)) \in B_2\}}$.

Il s'agit de la même proportion que celle (15) ci-dessus.

γ- $\mathcal{P}\{\xi_i(x)\xi_i(y)\xi_i(z)\xi_i(t)/((x,y),(z,t))\in G\}$.

C'est la proportion de parties X_i de cardinal $n(i)$ qui incluent les quatre objets x, y, z et t ; elle est égale à

$$\frac{n(i)[n(i)-1][n(i)-2][n(i)-3]}{n(n-1)(n-2)(n-3)} . \qquad (16)$$

Ainsi l'espérance mathématique de l'expression en titre de ce sous paragraphe 2.4.1. est égale à

$$\begin{aligned} & n(i)[n(i)-1]\, n(j)\, n(h)/ n(n-1) \\ + \; & n(i)[n(i)-1][n(i)-2]\, n(j)\, n(h)[n(h)-1]/n(n-1)(n-2) \\ + \; & n(i)[n(i)-1][n(i)-2]\, n(j)[n(j)-1]\, n(h)/n(n-1)(n-2) \\ + \; & n(i)[n(i)-1][n(i)-2][n(i)-3]\, n(j)[n(j)-1]\, n(h)[n(h)-1]/ \\ & \qquad n(n-1)(n-2)(n-3). \end{aligned} \qquad (17)$$

2.4.2- Calcul de l'espérance mathématique de

$$\left(\mathcal{E}\{\xi_i(x)\xi_i(y)/(x,y)\in F_j\times F_h\}\right)\left(\mathcal{E}\{\xi_{i'}(x')\xi_{i'}(y')/(x',y')\in F_j\times F_h\}\right).$$

Nous allons effectuer des calculs parallèles à ceux du paragraphe 2.4.1. ci-dessus. Dans ces conditions l'expression ci-dessus se décompose comme suit :

$$\begin{aligned} & \mathcal{E}\{\xi_i(x)\xi_i(y)\xi_{i'}(x)\xi_{i'}(y) / (x,y)\in F_j\times F_h\} \\ + & \mathcal{E}\{\xi_i(x)\xi_i(y)\xi_{i'}(x)\xi_{i'}(t)/((x,y),(x,t))\in B_1\} \\ + & \mathcal{E}\{\xi_i(x)\xi_i(y)\xi_{i'}(z)\xi_{i'}(y)/((x,y),(z,y))\in B_2\} \\ + & \mathcal{E}\{\xi_i(x)\xi_i(y)\xi_{i'}(z)\xi_{i'}(t)/((x,y),(z,t))\in G\} . \end{aligned} \qquad (18)$$

Puisque i est distinct de i', chacune des trois premières sommes est nulle. Il nous reste à évaluer l'espérance mathématique de l'élément courant de la dernière somme.

$\mathcal{E}\{\xi_i(x)\xi_i(y)\xi_{i'}(z)\xi_{i'}(t) / ((x,y),(z,t)) \in C\}$ représente la proportion de couples de parties $(X_i, X_{i'})$ telles que X_i (resp. $X_{i'}$) renferme les objets x et y (resp. z et t). Elle est égale à :

$$\frac{n(i)[n(i)-1]\,n(i')[n(i')-1]}{n(n-1)(n-2)(n-3)} \qquad (19)$$

Ainsi, l'espérance mathématique de l'expression en titre du paragraphe 2.4.2. est égale à

$$\frac{n(i)[n(i)-1]\,n(i')[n(i')-1]\,n(j)[n(j)-1]\,n(h)[n(h)-1]}{n(n-1)(n-2)(n-3)}. \qquad (20)$$

2.4.3.– Expression de la variance de la v.a. $S(j,h/P)$.

Compte tenu des expressions (11), (17) et (20), ainsi que celle (9), on obtient l'écriture détaillée suivante de la variance :

$$\boxed{\begin{aligned}
&\sum_{1\leq i\leq |I|} \frac{[n(i)-1]}{n(i)} \times \frac{n(j)\,n(h)}{n(n-1)}\left\{1+\frac{[n(i)-2][n(h)-1]}{(n-2)}\right.\\
&\left.+\frac{[n(i)-2][n(j)-1]}{(n-2)}+\frac{[n(i)-2][n(i)-3][n(j)-1][n(h)-1]}{(n-2)(n-3)}\right\}\\
&+2\sum_{1\leq i<i'\leq |I|} \frac{[n(i)-1][n(i')-1]\,n(j)[n(j)-1]\,n(h)[n(h)-1]}{n(n-1)(n-2)(n-3)}\\
&-\left[\frac{(n-|I|)}{(n-1)}\right]^2 \times \left(\frac{n(j)\,n(h)}{n}\right)^2 .
\end{aligned}} \qquad (21)$$

Compte tenu des formules (9) et (21), il résulte l'expression de l'indice centré et réduit.

2.5 – Forme duale de l'hypothèse d'absence de lien.

Nous allons ici fixer la partition $\{E_i / 1 \leq i \leq |I|\}$ et associer à la partition $\{F_j / 1 \leq j \leq J\}$, une partition aléatoire $\{Y_j / 1 \leq j \leq |J|\}$, en classes étiquetées et de type $s = [m(1), m(2), \ldots, m(j), \ldots, m(|J|)]$, dans l'ensemble $\mathcal{P}(n; s)$, muni d'une probabilité uniformément répartie, des partitions de E, en classes étiquetées de type s.

Reprenons l'expression de l'indice brut (5) (§ 2.1 ci-dessus):

$$s(j,h/P) = \mathcal{E}\left\{ \operatorname{card}(E_i \cap F_j) \operatorname{card}(E_i \cap F_h)/n(i) \,\middle|\, 1 \leq i \leq |I| \right\}. \quad (21)$$

Si la v.a. considérée ci-dessus se met sous la forme

$$S(j,h/P) = \mathcal{E}\left\{ \operatorname{card}(X_i \cap F_j) \operatorname{card}(X_i \cap F_h)/n(i) \,\middle|\, 1 \leq i \leq |I| \right\}, \quad (22)$$

celle que nous envisageons ici s'écrit

$$T(j,h/P) = \mathcal{E}\left\{ \operatorname{card}(E_i \cap Y_j) \operatorname{card}(E_i \cap Y_h)/n(i) \,\middle|\, 1 \leq i \leq |I| \right\}. \quad (23)$$

En désignant par ε_i (resp. η_j) la fonction indicatrice de E_i (resp. Y_j), $1 \leq i \leq |I|$ (resp. $1 \leq j \leq |J|$), la dernière v.a. se met sous la forme

$$T(j,h/P) = \mathcal{E}\left\{ \frac{1}{n(i)} \left[\mathcal{E}_{x \in E}\, \varepsilon_i(x)\, \eta_j(x)\right] \left[\mathcal{E}_{y \in E}\, \varepsilon_i(y)\, \eta_h(y)\right] \,\middle|\, 1 \leq i \leq |I| \right\}. \quad (24)$$

Nous allons nous rendre compte que cette dernière v.a. a la même distribution que $S(j,h/P)$.

<u>Propriété</u>. La distribution de la suite de v.a. entières $\{\text{card}\,(X_i \cap F_j) / 1 \leq j \leq |J|\}$ est identique à celle de la suite $\{\text{card}(E_i \cap Y_j) / 1 \leq j \leq |J|\}$.

En effet, $\Pr\{\text{card}(X_i \cap F_j) = n(i \wedge j) / 1 \leq j \leq |J|\}$ est définie par la proportion de parties dans $\mathcal{P}_{n(i)}(E)$ (ensemble des parties de E de même cardinal $n(i)$) pour lesquels la suite des cardinaux des intersections avec la suite des classes F_j est $\{n(i \wedge j) / 1 \leq j \leq |J|\}$. Elle est par conséquent égale à

$$\frac{\binom{m(1)}{n(i\wedge 1)}\binom{m(2)}{n(i\wedge 2)}\cdots\binom{m(j)}{n(i\wedge j)}\cdots\binom{m(|J|)}{n(i\wedge |J|)}}{\binom{n}{n(i)}}. \qquad (26)$$

Après développement et reorganisation, on peut ramener l'expression précédente à la forme suivante :

$$\frac{\dfrac{n(i)!}{n(i\wedge 1)!\,n(i\wedge 2)!\cdots\cdots n(i\wedge |J|)!}\times\dfrac{n(\bar{i})!}{n(\bar{i}\wedge 1)!\,n(\bar{i}\wedge 2)!\cdots n(\bar{i}\wedge |J|)!}}{\dfrac{n!}{m(1)!\,m(2)!\cdots m(j)!\cdots m(|J|)!}}, \qquad (26)$$

où nous avons noté $n(\bar{i}) = [n - n(i)] = n(E - E(i)) = n(E(\bar{i}))$, $n(\bar{i} \wedge j) = n(E(\bar{i}) \cap F_j)$, $1 \leq j \leq |J|$.

Or le dernier rapport (26) représente la proportion de partitions dans $\mathcal{P}(n;s)$ pour lesquelles la j-ème classe contient $n(i \wedge j)$ éléments de E_i, $1 \leq j \leq |J|$. Il s'agit donc de

$\Pr\{\text{card}(E_i \cap Y_j) = n(i \wedge j) / 1 \leq j \leq |J|\}$.

D'autre part, le système des relations stochastiques entre les différentes v.a. $\mathrm{card}(X_i \cap F_j)$, $1 \leq j \leq |J|$, est le même que celui entre les différentes v.a. $\mathrm{card}(E_i \cap Y_j)$, $1 \leq j \leq |J|$. Dans ces conditions, nous avons la propriété suivante de dualité:

<u>Théorème.</u> Les v.a. $S(j,h/P)$ et $T(j,h/P)$ (cf. formules (22) et (23)) ont la même distribution.

Il est d'ailleurs intéressant de calculer directement la moyenne et la variance de la v.a. $T(j,h/P)$ (cf. formule (24)) et de retrouver les formules (9) et (21) ci-dessus.

Si les $n(i)$, $n(j)$ et $n(h)$ sont, comme c'est le cas courant de tables de contingence, assez grands, on a la valeur très approchée suivante de la variance:

$$\sum_{1 \leq i \leq |I|} \frac{n(j)n(h)}{n} \left\{ \frac{1}{n} + \frac{n(i)n(h)}{n} + \frac{n(i)n(j)}{n} + \frac{[n(i)]^2 n(j)n(h)}{n^2} \right\}$$

$$+ 2 \sum_{1 \leq i < i' \leq |I|} \left[\frac{n(j)n(h)}{n} \right]^2 \frac{n(i)n(i')}{n^2}$$

$$- \left[1 - \frac{|I|}{n} \right]^2 \left[\frac{n(j)n(h)}{n} \right]^2$$

$$= n\,p(j)p(h) \left\{ \frac{|I|}{n} + [p(j)+p(h)] + n\,p(j)p(h) \sum_i [p(i)]^2 \right\}$$

$$+ 2[n\,p(j)p(h)]^2 \sum_{i<i'} p(i)p(i') - \left[1 - \frac{|I|}{n} \right]^2 [n\,p(j)p(h)]^2$$

$$= e(j,h) \left\{ \frac{|I|}{n} + [p(j)+p(h)] \right\} + e^2(j,h) \times \frac{|I|}{n} \left(2 - \frac{|I|}{n} \right), \quad (27)$$

où nous avons noté $e(j,h) = n\,p(j)p(h)$.

Il s'agit d'une expression de nature différente que

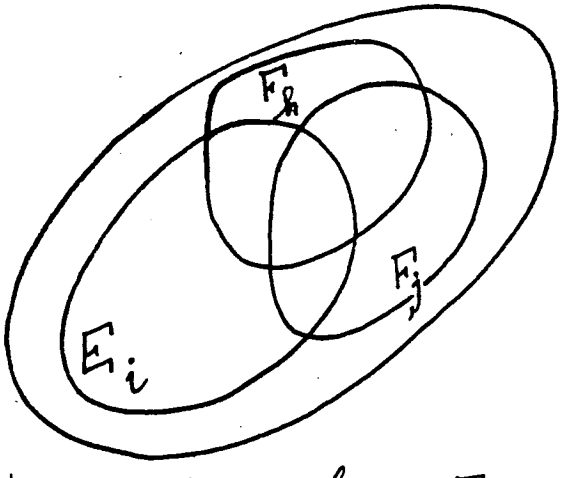
celle définie par le dénominateur élevé au carré de la formule (3) du paragraphe 1 ci-dessus.

III.2 CAS D'UNE JUXTAPOSITION HORIZONTALE DE TABLES DE CONTINGENCE.

La structure de la donnée est définie au paragraphe I ci-dessus (cf. expressions (1) et (2) § I) où il s'agit de comparer deux modalités j et h non nécessairement exclusives ; c'est à dire, telles que $j \in J^{(l)}$, $h \in J^{(m)}$, où l peut ou non être distinct de m, $1 \leq l, m \leq L$.

1 - L'indice brut et l'hypothèse d'absence de lien.

La situation peut être shématisée comme suit, relati-



vement à une même classe E_i. L'indice brut de proximité est le même que dans le cas précédent où F_j et F_h étaient disjoints (cf. formule (5) du paragraphe 2.1 de III.1.). Rappelons ici sa forme :

$$\left\{ \frac{n(i \wedge j)\, n(i \wedge h)}{n(i)} \Big/ i \in I \right\}. \quad (1)$$

L'hypothèse d'absence de lien va correspondre à la forme adoptée au paragraphe 2.5. ci-dessus, en ayant un caractère plus libre.

De façon précise, on fixe la partition $\{E_i / 1 \leqslant i \leqslant |I|\}$ et on associe au couple de parties (F_j, F_h), un couple de parties aléatoires indépendantes $(\mathcal{Y}_j, \mathcal{Y}_h)$ où $\mathcal{Y}_j$ (resp. $\mathcal{Y}_h$) est un élément aléatoire dans l'ensemble, muni d'une probabilité uniformément répartie, des parties de E de même cardinal $n(j) = \text{card}(F_j)$ (resp. $n(h) = \text{card}(F_h)$).

La v.a. associée à l'indice brut (1) se met dans ces conditions sous la forme

$$U(j,h/P) = \sum_{1 \leqslant i \leqslant |I|} \{\text{card}[E(i) \cap \mathcal{Y}_j] \, \text{card}[E(i) \cap \mathcal{Y}_h] / n(i)\}. \qquad (2)$$

Nous désignerons par $\nu(i \wedge j)$ (resp. $\nu(i \wedge h)$) la v.a. $\text{card}[E(i) \cap \mathcal{Y}_j]$ (resp. $\text{card}[E(i) \cap \mathcal{Y}_h]$), laquelle est hypergéométrique de paramètres $[n, n(i), n(j)]$ (resp. $[n, n(i), n(h)]$). De plus, les deux v.a. $\nu(i \wedge j)$ et $\nu(i \wedge h)$ sont indépendantes, compte tenu de l'indépendance de $\mathcal{Y}_j$ et de $\mathcal{Y}_h$.

2 – <u>Moyenne et Variance de la v.a. $U(j,h/P)$.</u>

L'indépendance entre $\nu(i \wedge j)$ et $\nu(i \wedge h)$ permet d'écrire

$$\mathcal{E}(U(j,h/P)) = \sum_{1 \leqslant i \leqslant |I|} \frac{1}{n(i)} \times \frac{n(i)\,n(j)}{n} \times \frac{n(i)\,n(h)}{n}$$

$$= \frac{n(j)\,n(h)}{n} \qquad (3)$$

On a d'autre part :

$$U^2(j,h/P) = \sum_{1 \leqslant i \leqslant |I|} \frac{\nu^2(i \wedge j)\,\nu^2(i \wedge h)}{n^2(i)}$$

$$+ 2 \sum_{1 \leqslant i < i' \leqslant |I|} \frac{\nu(i \wedge j)\,\nu(i \wedge h)}{n(i)} \times \frac{\nu(i' \wedge j)\,\nu(i' \wedge h)}{n(i')} \qquad (4)$$

Compte tenu des relations suivantes

$$\left.\begin{aligned}
\mathcal{E}[\nu^2(i\wedge j)] &= \frac{n(i)[n(i)-1]\,n(j)[n(j)-1]}{n(n-1)} + \frac{n(i)\,n(j)}{n}\\
\mathcal{E}[\nu^2(i\wedge h)] &= \frac{n(i)[n(i)-1]\,n(h)[n(h)-1]}{n(n-1)} + \frac{n(i)\,n(h)}{n}\\
\mathcal{E}[\nu(i\wedge j)\,\nu(i'\wedge j)] &= \frac{n(i)\,n(i')\,n(j)[n(j)-1]}{n(n-1)}\\
\mathcal{E}[\nu(i\wedge h)\,\nu(i'\wedge h)] &= \frac{n(i)\,n(i')\,n(h)[n(h)-1]}{n(n-1)}
\end{aligned}\right\} \qquad (5)$$

que le lecteur cherchera à retrouver, on a

$$\begin{aligned}
\mathcal{E}[U^2(j,h/P)] = {} & \sum_{1\leq i\leq |I|} \frac{1}{n(i)^2}\left[\frac{n(i)[n(i)-1]\,n(j)[n(j)-1]}{n(n-1)} + \frac{n(i)\,n(j)}{n}\right]\\
& \times\left[\frac{n(i)[n(i)-1]\,n(h)[n(h)-1]}{n(n-1)} + \frac{n(i)\,n(h)}{n}\right]\\
& + 2\sum_{1\leq i<i'\leq |I|} \frac{1}{n(i)\,n(i')}\left[\frac{n(i)\,n(i')\,n(j)[n(j)-1]}{n(n-1)}\right]\\
& \times\left[\frac{n(i)\,n(i')\,n(h)[n(h)-1]}{n(n-1)}\right].
\end{aligned} \qquad (6)$$

De sorte que l'expression de la variance de $U(j;h/P)$ peut se mettre sous la forme suivante :

$$\mathrm{var}[U(j,h/P)] = e(j,h)\Big\{\sum_{1\leq i\leq |I|}\Big\{\frac{[n(i)-1]^2[n(j)-1][n(h)-1]}{n(n-1)^2}$$

$$+\frac{[n(i)-1][n(j)-1]}{n(n-1)}+\frac{[n(i)-1][n(h)-1]}{n(n-1)}+\frac{1}{n}\Big\}$$

$$+2\sum_{1\leq i<i'\leq |I|}\frac{n(i)\,n(i')[n(j)-1][n(h)-1]}{[n(n-1)]^2}-e(j,h)\Big\}, \quad (7)$$

où, rappelons le, nous notons $e(j,h)$ le paramètre $n(j)n(h)/n$.

Considérons à présent la valeur approximative de (7) dans le cas où $n(j)$, $n(h)$ et $n(i)$ sont «assez» grands

$$\mathrm{var}[U(j,h/P)] = e(j,h)\Big\{e(j,h)+[p(j)+p(h)]+\frac{|I|}{n}-e(j,h)\Big\}$$

$$= e(j,h)\Big\{p(j)+p(h)+\frac{|I|}{n}\Big\} \quad . \quad (8)$$

On notera la différence avec le cas où j et h sont exclusifs ; pour la moyenne, il y a lieu de comparer l'expression (3) ci-dessus avec celle (9) du paragraphe III.1-2.2. et pour la variance, on comparera plus aisément la valeur (8) ci-dessus avec celle (27) du paragraphe précédent.

Comme on a pu déjà le constater, l'indice centré est exactement, au coefficient n près, le numérateur de l'indice de corrélation $\rho(j,h)$ défini par la formule (3) du paragraphe III.1-1. ci-dessus.

Nous allons à présent examiner ce que devient l'indice centré réduit

$$\frac{s(j,h/P)-\mathcal{E}[U(j,h/P)]}{\sqrt{\mathrm{var}[U(j,h/P)]}}, \quad (9)$$

dans le cas où la partition $\{E(i)/i\in I\}$ est la plus fine où chaque classe contient exactement un objet.

Dans cette situation $n(i)=1$ pour tout $i=1,2,\ldots,n$ et $n(i\wedge j)\,n(i\wedge h)$ n'est égal à 1 que si et seulement si l'objet codé i possède les deux modalités-attributs j et h ; de sorte que

$$s(j,h/P) = n(j\wedge h) = \mathrm{card}\,(F_j \cap F_h)\,. \qquad (10)$$

D'autre part, la variance de $U(j,h/P)$ devient dans ce cas particulier

$$\mathrm{var}\,[U(j,h/P)] = e(j,h)\left\{1 + \frac{[n(j)-1][n(h)-1]}{(n-1)} - e(j,h)\right\}$$

$$= \frac{n(j)\,n(\bar{j})\,n(h)\,n(\bar{h})}{n^2\,(n-1)} \qquad (11)$$

où nous notons $n(\bar{j}) = [n - n(j)]$ (resp. $n(\bar{h}) = [n - n(h)]$)

On retrouve ainsi, au coefficient $\sqrt{n-1}$ près, l'indice d'association de K. Pearson entre les deux attributs-modalités j et h.

Pour cette même situation où la partition $\{E(i)/i\in I\}$ est discrète, nous allons examiner ce que devient l'indice de corrélation $\rho(j,h)$ défini par la formule (3) du paragraphe III.1.1.

Cet examen peut être conçu relativement à un tableau disjonctif complet croisant l'ensemble I des individus ou objets avec la suite des modalités de c variables qualitatives nominales. On a

$$\rho(j,h) = \frac{\sum_{1\leq i\leq n} \frac{1}{n}\left(\frac{\varepsilon(i\wedge j)}{c} - \frac{p(j)}{c}\right)\left(\frac{\varepsilon(i\wedge h)}{c} - \frac{p(h)}{c}\right)}{\left\{\left[\sum_{1\leq i\leq n} \frac{1}{n}\left(\frac{\varepsilon(i\wedge j)}{c} - \frac{p(j)}{c}\right)^2\right]\left[\sum_{1\leq i\leq n} \frac{1}{n}\left(\frac{\varepsilon(i\wedge h)}{c} - \frac{p(h)}{c}\right)^2\right]\right\}^{\frac{1}{2}}}, \qquad (12)$$

où $\varepsilon(i\wedge j)$ (resp. $\varepsilon(i\wedge h)$) est égal 1 ou 0 selon que l'attribut j (resp. h) est présent ou absent chez le sujet i et où $p(j)$ (resp. $p(h)$) est égal à $n(j)/n$ (resp. $n(h)/n$).

L'indice (12) se réduit exactement à celui de K. Pearson.

<u>Théorème.</u> Dans le cadre de l'h.a.l. définie ci-dessus (§ III.2), l'indice $s(j,h/P)$ centré (numérateur de (9)) est, au coefficient n près, égal au numérateur de l'indice de corrélation $\rho(j,h)$. Le dénominateur de l'indice centré réduit (9) reste de forme essentiellement différente du dénominateur de $\rho(j,h)$. Toutefois, les deux indices coïncident dans le cas d'un tableau disjonctif complet où la partition $\{E(i)/i \in I\}$ est discrète.

Pour établir la table des indices d'association entre éléments de $J = J^{(1)} \cup J^{(2)} \cup \cdots \cup J^{(\ell)} \cup \cdots \cup J^{(L)}$ dans le cas d'une juxtaposition horizontale de tableaux de contingence, conformément au point de vue développé, on peut se contenter d'adopter la dernière forme de l'h.a.l. que les modalités à comparer soient non-exclusives ou exclusives. Mais, rien n'empêche d'être plus précis et d'adopter la formule (9) ci-dessus de comparaison si les deux modalités j et h ne sont pas nécessairement exclusives et celle, résultant de l'h.a.l. du paragraphe III.1, si les modalités j et h sont exclusives (i.e. appartiennent à un même $J^{(\ell)}$).

IV. SUR LA CONSTRUCTION ASCENDANTE D'UN ARBRE BINAIRE DES CLASSIFICATIONS SUR J.

B. Tallur [TALLUR (1982)] utilise l'indice de corrélation $\rho(j,h)$ pour former la totalité de l'arbre binaire des classifications sur J selon l'algorithme classique de classification ascendante hiérarchique. En effet, la fusion des deux modalités j et h les plus voisines conduit à la création d'une nouvelle modalité $j \vee h$ (j ou h) et on se retrouve à chaque étape de l'algorithme à rechercher dans l'ensemble des nouvelles modalités les deux les plus proches.

Certes, il peut se faire qu'à un même niveau de la construction, on trouve plusieurs paires de modalités également les plus

voisines ; dans ce cas, on procède — dans un ordre quelconque — à la suite des agrégations des paires les plus proches et les classes résultantes sont placées à un même niveau de l'arbre.

Pour que l'arbre soit sans inversions ; c'est à dire, pour que la suite des valeurs de l'indice de proximité entre les deux classes les plus proches à un niveau donné de l'arbre, B. Tallur montre [TALLUR (1982)] qu'il importe d'apporter un cœfficient correctif et d'adopter comme indice d'association entre la nouvelle agrégation $j \vee h$ et l $(l \neq j,\ l \neq h)$ $\frac{1}{\sqrt{2}}\, \rho(j \vee h, l)$.

Un tel cœfficient $(1/\sqrt{2})$ qu'il importe de justifier de façon plus précise, peut très intuitivement se comprendre si on veut préserver le caractère entier des unités statistiques initiales que sont les modalités de J. De toute façon de la sorte, l'algorithme donne d'excellents résultats.

L'indice (9) auquel nous sommes parvenus ci-dessus, ne diffère de $\rho(j,h)$ qu'au niveau du dénominateur ; son expression explicite est la suivante :

$$Q(j,h) = \frac{\sum_{1 \leq i \leq n} \frac{n(i \wedge j)\, n(i \wedge h)}{n(i)} - \frac{n(j)\, n(h)}{n}}{\sqrt{n\, p(j)\, p(h)\, [p(j) + p(h) + \iota]}}, \qquad (1)$$

où $\iota = |I|/n$.

L'objet de ce bref paragraphe est précisément de montrer que le même cœfficient multiplicatif $(1/\sqrt{2})$ s'impose dans une formation ascendante et sans inversions de l'arbre des classifications, basée sur le seul indice $Q(j,h)$.

Il s'agit de déterminer le cœfficient positif α tel que

$$Q(j,l) < Q(j,h) \text{ et } Q(h,l) < Q(j,h) \Longrightarrow \alpha\, Q(j \vee h, l) < Q(j,h). \quad (2)$$

La première inégalité $Q(j,l) < Q(j,h)$ se met sous la forme

$$\sum_i \frac{n(i \wedge j)\, n(i \wedge l)}{n(i)} - \frac{n(j)\, n(l)}{n} < \sqrt{n\, p(j)\, p(l)\, [p(j) + p(l) + \iota]}\; Q(j,h), \quad (3)$$

la deuxième inégalité $Q(h,l) < Q(j;h)$ se met également sous la forme

$$\sum_i \frac{n(i \wedge h)\, n(i \wedge l)}{n(i)} - \frac{n(h)\, n(l)}{n} < \sqrt{n\, p(h)\, p(l)\, [p(h) + p(l) + 1]}\; Q(j;h). \quad (4)$$

L'addition membre à membre des inégalités (3) et (4) conduit à l'inégalité qui est obtenue dans le cas où j et h sont exclusives

$$\text{num.}\,[Q(j \vee h, l)] < \sqrt{n\, p(l)}\, \Big\{ \sqrt{p(j)\,[p(j) + p(l) + 1]} + \sqrt{p(h)\,[p(h) + p(l) + 1]} \Big\}\, Q(j;h), \quad (5)$$

où num. désigne le numérateur. Il en résulte que

$$Q(j \vee h, l) < \frac{\sqrt{p(j)\,[p(j) + p(l) + 1]} + \sqrt{p(h)\,[p(h) + p(l) + 1]}}{\sqrt{[p(j) + p(h)]\,[p(j) + p(h) + p(l) + 1]}}\; Q(j;h). \quad (6)$$

En posant $a = p(j)\,[p(j) + p(l) + 1]$ et $b = p(h)\,[p(h) + p(l) + 1]$, on a a fortiori

$$Q(j \vee h, l) < \frac{\sqrt{a} + \sqrt{b}}{\sqrt{a + b}}\; Q(j;h). \quad (7)$$

Or, on peut facilement voir que, a et b étant deux nombres réels positifs

$$\frac{\sqrt{a} + \sqrt{b}}{\sqrt{a + b}} < \sqrt{2}, \quad (8)$$

d'où, le résultat annoncé que nous venons d'établir dans le cas où J est formé de modalités exclusives et qui, en toute rigueur, concerne l'agrégation des deux premières modalités les plus proches.

De toute façon, que les modalités de J soient exclusives (une seule table de contingence) ou non (juxtaposition horizontale de tables de contingence), l'algorithme procède de la même façon par addition de colonnes, de sorte que reste vraie la propriété qu'on vient de mentionner et qu'il convient d'exprimer comme suit :

$$\left.\begin{array}{l} Q(j,l) < Q(j,h) \\ Q(h,l) < Q(j,h) \end{array}\right\} \Longrightarrow \frac{1}{\sqrt{2}}\, Q(j+h,l) < Q(j,h) \quad . \qquad (9)$$

Nous allons maintenant pour achever notre démonstration donner l'expression explicite de l'indice d'association entre deux classes de modalités de J et montrer que, pour cet indice, l'arbre total est sans inversions.

Si G et H sont deux classes disjointes de modalités de J regroupant respectivement l et m éléments de J, l'indice d'association entre G et H, $Q_c(G,H)$ est défini par l'expression suivante

$$\left.\begin{array}{l} Q_c(G,H) = \left(\frac{1}{\sqrt{2}}\right)^{l+m-1} Q(G,H) \\ \text{où} \quad Q(G,H) = Q\left(\sum\{n(i\wedge g)/g\in G\}, \sum\{n(i,h)/h\in H\}\right) \end{array}\right\} \qquad (10)$$

Dans ces conditions, nous avons à prouver que

$$\left.\begin{array}{l} Q_c(G,F) < Q_c(G,H) \\ Q_c(H,F) < Q_c(G,H) \end{array}\right\} \Longrightarrow Q_c(G+H,F) < Q_c(G,H) \ , \qquad (11)$$

où on suppose que F est une classe formée de k modalités de J, disjointe de celles G et H.

Compte tenu de (10), les relations du premier membre de (11) s'expriment par

$$\left.\begin{array}{l} Q(G,F) < \left(\frac{1}{\sqrt{2}}\right)^{m-k} Q(G,H) \\ Q(H,F) < \left(\frac{1}{\sqrt{2}}\right)^{l-k} Q(G,H) \end{array}\right\} \qquad (12)$$

Il en résulte, avec des notations que l'on comprend, les inégalités suivantes analogues à celles (3) et (4) ci-dessus :

$$\left.\begin{array}{l} \text{num.}\, Q(G,F) < \sqrt{n\, p(G)\, p(F)[p(G)+p(F)+L]}\left(\frac{1}{\sqrt{2}}\right)^{m-k} Q(G,H) \\ \text{num.}\, Q(H,F) < \sqrt{n\, p(H)\, p(F)[p(H)+p(F)+L]}\left(\frac{1}{\sqrt{2}}\right)^{l-k} Q(G,H) \end{array}\right\} \quad (13)$$

Il en résulte qu'en notant

$A = p(G)[p(G)+p(F)+L]$, $B = p(H)[p(H)+p(F)+L]$,

on ait a fortiori, de la même façon que pour la relation (7) ci-dessus,

$$\left(\frac{1}{\sqrt{2}}\right)^{k} Q(G+H,F) < \frac{\sqrt{A}\left(\frac{1}{\sqrt{2}}\right)^{m} + \sqrt{B}\left(\frac{1}{\sqrt{2}}\right)^{l}}{\sqrt{A+B}}\, Q(G,H). \quad (14)$$

Puisque $\min(l,m) \geq 1$, le coefficient de $Q(G,H)$ dans le second membre de (14) est, compte tenu de la relation (8) ci-dessus, strictement inférieur à 1. Donc, en multipliant les deux membre de (14) par $(1/\sqrt{2})^{l+m-1}$, on obtient le résultat annoncé dans le second membre de (11).

<u>Théorème</u>. L'indice d'association entre deux classes disjointes G et H de modalités de J ($l = \text{card}(G)$, $m = \text{card}(H)$) qui généralise l'indice $Q(j,h)$ de la formule (1) ci-dessus, pour l'obtention par l'algorithme de classification hiérarchique ascendante d'un arbre binaire sans inversions est défini par

$$Q_c(G,H) = \left(\frac{1}{\sqrt{2}}\right)^{l+m-1} Q(G,H)$$

où $Q(G,H)$ est, rappelons le, l'indice (1) ci-dessus appliqué à deux colonnes sommes : la première (resp. la seconde) résultant de la somme des colonnes des modalités initiales de G (resp. de H).

IV - CONCLUSION ; QUELQUES EXTENSIONS.

Une extension naturelle de l'approche pour définir un indice d'association concerne également le cas d'un tableau de données Individus x Variables quantitatives où l'unité utilisée est commune aux différentes variables (franc, kilo, unité de surface ou de volume, ...). Un exemple typique est celui où on cherche à étudier la répartition des dépenses, sur un ensemble de postes, de l'ensemble des ménages — dont on dispose d'un échantillon représentatif — d'une région économique donnée, pendant une période fixée.

On peut en effet dans ce cas donner une interprétation très claire du lien "brut" entre deux ménages sur un poste de dépense fixé, respectivement, de deux postes de dépense relativement à un ménage donné. Finalement, tout se passe comme pour un tableau de contingence qui définit le croisement de deux partitions — la première par ménage et la seconde par poste — sur la masse monétaire dépensée par l'échantillon observé sur l'ensemble considéré des postes de dépense.

La deuxième extension que nous voulons pour terminer évoquer, concerne le passage de la notion d'association totale à celle, partielle et le sens qu'il faut donner à cette dernière.

Si on considère le point de vue géométrique qui a prévalu à la définition du coefficient d'association $\rho(j, j')$ (cf. formule (3') § III.1.1.), on peut poursuivre l'analogie formelle avec le cas linéaire pour obtenir le coefficient d'association partielle $\rho(j, j' ; h)$ qui doit neutraliser l'influence de la modalité h dans la comparaison de j à j'.

u, v et w étant trois variables numériques de description d'un ensemble I d'individus, le coefficient de corrélation partielle $\rho(v, w ; u)$ qui est le coefficient de corrélation entre les résidus $[v - v(u)]$ et $[w - w(u)]$, apparait comme le

coefficient de corrélation entre les deux variables :

$$\left[\frac{(v-\bar{v})}{s(v)} - \text{cor.}(u,v) \times \frac{(u-\bar{u})}{s(u)}\right].$$

et

$$\left[\frac{(w-\bar{w})}{s(w)} - \text{cor.}(u,w) \times \frac{(u-\bar{u})}{s(u)}\right] . \qquad (1)$$

Partant de là, on obtiendra par identification formelle, moyennant la représentation géométrique du paragraphe III.1.1., le coefficient d'association partielle relativement à la modalité h, entre les deux modalités j et j' :

$$\rho(j,j';h) = \frac{\rho(j,j') - \rho(j,h)\rho(j',h)}{\sqrt{1-\rho^2(j,h)}\,\sqrt{1-\rho^2(j',h)}} . \qquad (2)$$

Nous avions bien défini dans [LERMAN (1981 b)] des coefficients d'association partielle entre variables qualitatives où on ne se réfère nullement à une représentation géométrique ou linéaire de la représentation des variables. Dans ces conditions, la question se pose de savoir si nous pouvons faire de même dans la situation étudiée ici.

BIBLIOGRAPHIE.

HÁJEK J., " Some extensions of the Wald - Wolfowitz - Noether theorem.", Annals of Mathematical Statistics 32, 506-523, (1961).

LERMAN I.C., " Classification et analyse ordinale des données.", Dunod, 760p., Paris (1981).

LERMAN I.C., " Corrélation partielle dans le cas « qualitatif »", rapport interne IRISA n° 153, 125p, Rennes (1981b). A paraître sous la forme de deux articles ; le premier dans les "Publications de l'Institut de Statistique de l'Université de Paris" et le second dans la R.A.I.R.O. série verte.

LERMAN I.C. et TALLUR B., " Classification des éléments constitutifs d'une juxtaposition de tableaux de contingence", Revue de Statistique Appliquée, vol. XXVIII, n°3, 1980.

NOETHER G.E., " On a theorem by Wald and Wolfowitz", Annals of Mathematical Statistics., vol. 20, p. 455, (1949).

TALLUR B., " Un nouvel algorithme de classification hiérarchique des éléments constitutifs d'une juxtaposition de tableaux de contingence, basé sur la corrélation", rapport interne IRISA n° 177, Rennes (1982).

TALLUR B., " Etude de l'agriculture régionale Française", rapport interne IRISA n° 103, Rennes (1978).

WALD A. and WOLFOWITZ J., " Statistical tests based on permutations of the observations.", Annals of Mathematical Statistics., vol. 15, p. 358, (1944)

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique