Pub. Inst. Stat. Univ.
XXIX, fasc. 3-4, 27-57

JUSTIFICATION ET VALIDITE STATISTIQUE D'UNE ECHELLE $[0,1]$ DE FREQUENCE MATHEMATIQUE POUR UNE STRUCTURE DE PROXIMITE SUR UN ENSEMBLE DE VARIABLES OBSERVEES

I.C. LERMAN
IRISA - Campus de Beaulieu - 35042 RENNES CEDEX

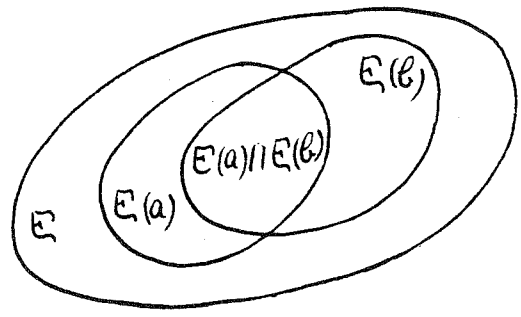
I. INTRODUCTION ET RAPPELS

V est un ensemble fini de m variables descriptives d'un même type, observées sur un ensemble E d'objets ou individus défini par un échantillon de la population étudiée $\mathcal{P}$. Notre but est ici de situer notre démarche dans l'élaboration d'une "mesure" d'association sur V, par rapport à un cadre d'inférence statistique classique où E se trouve regardé comme résultant d'un échantillonnage aléatoire dans $\mathcal{P}$, supposée de très grande taille. n et $\mathcal{N}$ désignant les cardinaux respectifs de E et $\mathcal{P}$.

Commençons par rappeler cette démarche -qui se situe au niveau de E sans aucune référence à $\mathcal{P}$- dans le cas le plus simple où l'ensemble des variables, qu'on notera dans ce cas $\mathcal{Q}$, est formé d'attributs logiques.

I. 1. ASSOCIATION ENTRE DEUX ATTRIBUTS ; Indice centré réduit, indice de la vraisemblance du lien .

Nous représentons un attribut par la partie de E formée des individus qui le possèdent ; de sorte que, relativement à un couple (a,b) d'attributs descriptifs, on peut représenter le diagramme suivant de Venn



où E (a) (resp. E (b)) est le sous-ensemble de E formé des individus qui possèdent l'attribut a (resp. b).

Le point de départ de la construction de l'indice d'association est l'indice "brut" : $s=card[E(a)\cap E(b)]$.

L'indice définitif "normalise" s en le situant par rapport à la distribution commune de l'une ou de l'autre des deux v.a. $S(a)=card[E(a)\cap Y]$ et $S(b)=car[X\cap E(b)]$ où X (resp. Y) est une partie aléatoire de E, de même cardinal n(a) que E(a) (resp. n(b) que E(b)) et prise uniformément au hasard ; en d'autres termes, X (resp. Y) est un élément aléatoire dans l'ensemble $\mathcal{P}_{n(a)}(E)$ (resp. $\mathcal{P}_{n(b)}(E)$) - muni d'une probabilité uniformément répartie - des parties de E de même cardinal n(a)(resp.n(b)).

Cette distribution est hypergéométrique de moyenne $\mu_{ab} = n(a)n(b)/n$ et de variance $\sigma^2_{ab} = n(a)n(\bar{a})n(b)n(\bar{b})/n^2(n-1)$, où $\bar{a}$ (resp. $\bar{b}$) désigne l'attribut opposé à a (resp. b) et où $n(\bar{a}) = card[E(\bar{a})] = n-n(a)$ (resp. $n(\bar{b}) = card[E(\bar{b})] = n-n(b)$). L'indice "centré réduit" $[s-\mu_{ab}]/\sigma_{ab}$ peut prendre la forme suivante

$$_1Q(a,b)=\sqrt{n-1}\ \ _1r(a,b), \quad (1)$$

où

$$_1r(a,b)=[p(a\wedge b)p(\bar{a}\wedge\bar{b})-p(a\wedge\bar{b})p(\bar{a}\wedge b)]/\sqrt{p(a)p(\bar{a})p(b)p(\bar{b})}, \quad (2)$$

et où p désigne la proportion définie au niveau de E ; ainsi, par exemple : $p(a\wedge b)=n(a\wedge b)/n=card[E(a)\cap E(b)]/card[E]$.

Q(a,b) n'est autre, au coefficient $\sqrt{(n-1)}$ près, que le coefficient d'association de K. Pearson. D'autre part, $Q(a,b)=Q(\bar{a},\bar{b})=-Q(a,\bar{b})=-Q(\bar{a},b)$, ce qui montre que l'indice obtenu aurait été le même si au lieu de partir de l'indice brut $s=card[E(a)\cap E(b)]$ (nombre d'associations "positives"), on partait de $t=card[E(\bar{a})\cap E(\bar{b})]$ (nombre d'associations "négatives"). Enfin, $Q^2(a,b)$ n'est autre que la statistique du χ^2 attachée au tableau de contingence 2x2 croisant les deux variables qualitatives dichotomiques ; les deux modalités de la première (resp. seconde) sont a et $\bar{a}$ (resp. b et $\bar{b}$).

Pour être conforme à nos précédentes notations (cf. par exemple [LERMAN(1981)]) désignons par N_1 l'hypothèse d'absence de liaison ci-dessus considérée et par S l'une ou l'autre des deux v.a. de même loi S(a) et S(b). L'indice de la vraisemblance du lien que nous avons introduit entre les deux attributs a et b seront considérés d'autant plus ressemblants que le nombre d'associations positives $s=card[E(a)\cap E(b)]$ est invraisemblablement grand, eu égard à la distribution de la v.a. S, que par conséquent $Pr\{S>s/N_1\}$ est petite ; c'est-à-dire, que $Pr\{S\leq s/N_1\}$ est grande. D'où l'idée de mesurer directement la "ressemblance" entre les attributs a et b par -ce que nous appelons- la "vraisemblance" P(a,b) qui représente une probabilité, ou -cf. la terminologie de M. Allais [M. ALLAIS(1983)]- une "fréquence mathématique":

$$P(a,b)=Pr\{S(a)\leq s/N_1\}=Pr\{S(b)\leq s/N_1\} \quad (3)$$

Cette probabilité a -dans le cadre de l'h.a.l. N_1- un sens très concret : en considérant la v.a. S(a)(resp.S(b)), il s'agit de la proportion dans $\mathcal{P}_{n(b)}(E)$ (resp. $\mathcal{P}_{n(a)}(E)$) de parties Y_1 (resp.X_1) dont l'intersection avec E(a)(resp.E(b)) réalise un cardinal inférieur ou égal à s.

Dans un test non paramétrique d'hypothèse -défini au niveau de E- on ne se sert de la probabilité $Pr\{S>s/N_1\}$ que pour rejeter ou non l'hypothèse d'absence de lien entre les deux attributs a et b. Alors que nous prétendons utiliser de façon beaucoup plus riche cette échelle de probabilité pour en faire véritablement une échelle de mesure de la liaison. Cette dernière ne prend tout son intérêt que lorsqu'il s'agit de comparer deux à deux, un ensemble de variables observées d'un même type (cf. § I.2. ci-dessous).

Cette approche est parfaitement conforme à l'optique de l'analyse des données qui est en quelque sorte opposée à celle des tests d'indépendance [LERMAN(1983)]. Si pour cette dernière, on privilégie la croyance en l'absence des liaisons entre les variables observées, au contraire, pour l'analyse des données, il n'y a aucun doute quant à l'existence de ces liaisons qui peuvent être plus ou moins fortes ou ténues et qu'il s'agit d'organiser au mieux.

Cette démarche n'est pas sans risque de glissement dans les notions et de précision dans les estimations calculées des ressemblances. L'objet de notre réflexion ici consiste précisément à montrer comment neutraliser ces deux risques et obtenir de la sorte une échelle très fine et très riche pour la comparaison deux à deux d'un ensemble de variables, ce qui conduit au critère très général de la "vraisemblance du lien" pour la construction ascendante hiérarchique d'un arbre des classifications sur l'ensemble des variables de description.

Nous avons pu mettre en évidence deux autres formes N_2 et N_3 de l'hypothèse d'absence de liaison ("h.a.l.") -entre deux attributs descriptifs- qui correspondent à deux modèles aléatoires des choix définis au niveau de l'ensemble des parties d'un ensemble [LERMAN(1981) Chap.2].

Pour la forme N_2 de l'h.a.l., on associe au couple (E(a),E(b)), un couple (X,Y) de parties aléatoires indépendantes de E, où le choix aléatoire de Y se fait selon un principe analogue à celui de X. Pour ce dernier choix, nous munissons l'ensemble $\mathcal{P}(E)$ des parties de E d'une mesure de probabilité plus diffuse que dans le cas de l'h.a.l. N_1 ; alors que pour le modèle aléatoire 1, la mesure de probabilité était concentrée sur le seul niveau du simplexe $\mathcal{P}(E)$ associé à $\mathcal{P}_{n(a)}(E)$, elle sera ici répartie sur les différents niveaux.

Ainsi, le modèle aléatoire 2 comporte deux pas : le premier consiste dans le choix d'un niveau et le second, dans le choix d'un élément de ce niveau. Pour le choix du niveau, on introduit la v.a. K indice d'un même niveau et cardinal commun de toutes les parties de ce niveau. K est considérée comme une v.a. binomiale de paramètres (n,p(a)) où p(a)=n(a)/n. Pour le choix aléatoire d'un élément de même niveau k, la probabilité binomiale affectée à ce niveau est uniformément répartie sur l'ensemble des $\binom{n}{k}$ sommets de ce niveau (dont chacun représente une partie de cardinal k).

Dans ces conditions, on montre que la v.a. card(X∩Y) suit une loi binomiale de paramètres $(n,\pi=p(a)p(b))$. Il en résulte que l'indice "centré réduit" associé à cette forme N_2 de l'h.a.l. peut se mettre sous la forme

$$ {}_2Q(a,b)=\sqrt{n}\ {}_2r(a,b) \qquad (4)$$

où

$$ {}_2r(a,b)=\left[p(a\wedge b)p(\bar{a}\wedge\bar{b})-p(a\wedge\bar{b})p(\bar{a}\wedge b)\right]\Big/\sqrt{p(a)p(b)\left[1-p_{(a)}p_{(b)}\right]} \qquad (5)$$

Le dénominateur du rapport qui définit ce dernier indice étant plus grand que celui de l'indice ${}_1r(a,b)$ (cf. formule (2) ci-dessus), on a également

$$-1\leq {}_2r(a,b)\leq 1. \qquad (6)$$

La forme N_3 de l'h.a.l. suppose un modèle aléatoire de choix de X(resp.de Y indépendamment de X) à trois pas. Par rapport aux deux modèles précédents, on associera ici à E un ensemble aléatoire $\mathcal{E}$, mais où le seul aléa qui nous intéresse concerne $\nu=\text{card}(\mathcal{E})$ qu'on suppose suivre une loi de Poisson de paramètre $n=\text{card}(E)$. Conditionnellement à $\mathcal{E}=E_o$, les deux pas suivants de ce modèle 3 sont définis de la même façon que pour le modèle binominal 2 ci-dessus. Ce modèle s'appelle Poissonnien car nous démontrons que la distribution de la v.a. $\text{card}(X\cap Y)$ est de Poisson de paramètre $n\pi$ où $\pi=p(a)p(b)$.

A cette forme N_3 de l'h.a.l., se trouve associé l'indice "centré réduit" qui s'écrit

$${}_3Q(a,b)=\sqrt{n}\ {}_3r(a,b), \qquad (7)$$

où on a

$${}_3r(a,b)=\left[p(a\wedge b)p(\bar{a}\wedge\bar{b})-p(a\wedge\bar{b})p(\bar{a}\wedge b)\right]\Big/\sqrt{p(a)p(b)} \qquad (8)$$

Le dénominateur de ce rapport étant plus grand que celui du rapport (5) définissant ${}_2r(a,b)$, on a a fortiori :

$$-1\leq {}_3r(a,b)\leq 1 \qquad (9)$$

On constate que les numérateurs des trois indices ${}_1r(a,b)$, ${}_2r(a,b)$ et ${}_3r(a,b)$ sont identiques. Si ce numérateur est positif, on a

$${}_3r(a,b)<{}_2r(a,b)<{}_1r(a,b) \qquad (10)$$

Le modèle Binomial fournit donc un indice "intermédiaire" entre les indices Poissonnien et Hypergéométrique. Si pour l'analyse statistique, le modèle Binomial est le plus souple à manipuler, par contre pour l'analyse des données, le choix se présente le plus clairement entre les indices ${}_3r(a,b)$ et ${}_1r(a,b)$.

L'indice ${}_1r(a,b)$ est parfaitement symétrique : si a et b sont deux attributs rares pour lesquels le numérateur commun des indices est positif, on a

$${}_1r(a,b)={}_1r(\bar{a},\bar{b})$$

où $\bar{a}$ et $\bar{b}$ sont les attributs respectivement opposés à a et b, alors que

$${}_3r(a,b)>{}_3r(\bar{a},\bar{b})$$

Le mathématicien épris de symétrie aura tendance à préférer l'indice r_1. Toutefois, en analyse des données où on travaille avec un ensemble $\mathcal{A}$ d'attributs "orientés" (i.e. $a \in \mathcal{A} \iff a \notin \mathcal{A}$), toutes choses "égales par ailleurs", il importe que l'association entre attributs rares soit plus ponctuée que celle entre attributs fréquents. C'est précisément ce que réalise le modèle N_3 de l'h.a.l.

Pour définir -relativement à l'association entre deux attributs a et b seulement- l'indice de la "vraisemblance du lien" $Pr\{S \leq s/N_\varepsilon\}$, on se sert de l'excellente approximation fournie par la loi normale $\mathcal{N}(0,1)$, de la distribution de $_\varepsilon Q(a',b')$, où a' et b' sont les deux attributs aléatoires respectivement associés à a et b dans le cadre de l'h.a.l. N_ε, ε=1,2 et 3. De sorte que

$$_\varepsilon P(a,b) = Pr\{S \leq s/N_\varepsilon\} \cong \Phi\left[\sqrt{n}\ {}_\varepsilon r(a,b)\right], \quad (11)$$

où Φ est la fonction de répartition (f.r.) $\mathcal{N}(0,1)$.

I.2. ASSOCIATIONS DEUX A DEUX SUR UN ENSEMBLE $\mathcal{A}$ D'ATTRIBUTS.

Pour n assez grand, on se rend compte et ce, d'autant mieux si on considère l'optique inférentielle, que P(a,b) atteint "facilement" 1 ou 0, selon que le signe de r(a,b) est positif ou négatif. Ainsi l'usage directe de la formule (11) pour établir la table des indices d'association entre éléments de $\mathcal{A}$ en termes de "vraisemblance du lien", conduit à une grande part de valeurs de ces indices par trop voisines de 1 ou de 0. C'est que le passage direct de $_\varepsilon Q(a,b)$ à l'échelle de fréquence mathématique $[0,1]$, au moyen de la transformation monotone $\Phi[{}_\varepsilon Q(a,b)]$, permet davantage la mise en évidence d'un lien fort entre les deux composantes d'une même paire d'attributs que l'évaluation comparée des liaisons entre les différentes paires d'un même ensemble $\mathcal{A}$ d'attributs ou de variables descriptives. On désignera par $P_2(\mathcal{A})$ l'ensemble des paires d'éléments de $\mathcal{A}$.

Nous avons pu nous rendre compte, qu'à des fins de comparaisons mutuelles entre éléments de $\mathcal{A}$, la statistique de proximité entre deux éléments donnés a et b de $\mathcal{A}$, doit avoir une nature intrinsèque à $\mathcal{A}$ et correspondre par conséquent à la contribution relative de l'association entre a et b, par rapport à l'ensemble des associations deux à deux entre éléments de $\mathcal{A}$. Une telle démarche est parfaitement en accord avec l'optique de l'analyse des données où ce qui est en question n'est pas tant de douter des liaisons que de les organiser au mieux les unes par rapport aux autres.

D'où la nécessité d'une réduction globale des similarités $\{Q(a,b)/\{a,b\} \in P_2(\mathcal{A})\}$ qui permettrait d'utiliser tout le pouvoir discriminant de l'échelle $[0,1]$ de probabilité de la f.r. de la loi normale $\mathcal{N}(0,1)$.

Après deux solutions expérimentalement étudiées [M.H. NICOLAU(1972), repris dans I.C. LERMAN(1973a), F. BONNIEUX, P. RAINELLI, T. CHANTREL et I.C.LERMAN (1977)], nous proposons ici une réduction globale au moyen du moment absolu d'ordre 2. En d'autres termes, on substituera à la table des indices

$$\{Q(a,b)/\{a,b\} \in P_2(\mathcal{A})\}, \quad (12)$$

celle

$$\{Q_s(a,b) = \frac{Q(a,b)}{\sqrt{M_2(Q)}} / \{a,b\} \in P_2(\mathcal{O})\}, \qquad (13)$$

où

$$M_2(Q) = \frac{1}{\binom{m}{2}} \sum \{Q^2(a,b) / \{a,b\} \in P_2(\mathcal{O})\}, \qquad (14)$$

le moment absolu d'ordre 2 de la distribution des similarités $\{Q(a,b)/\{a,b\} \in P_2(\mathcal{O})\}$.

D'un point de vue formel, nous remarquerons que l'indice $Q_s(a,b)$ qui rapporte $Q(a,b)$ à $\sqrt{M_2(Q)}$ se présente sous la forme d'une "densité orientée" en $\{a,b\}$ de $P_2(\mathcal{O})$ et qu'on a

$$\sum\{Q_s^2(a,b)/\{a,b\} \in P_2(\mathcal{O})\} = \binom{m}{2} \qquad (15)$$

L'important est la justification statistique de la table des indices de vraisemblance des liens, obtenue comme suit :

$$\{P(a,b) = \Phi[Q_s(a,b)]/\{a,b\} \in P_2(\mathcal{O})\}, \qquad (16)$$

où Φ est la f.r. de la loi $\mathcal{N}(0,1)$.

On tentera d'abord cette justification de façon non paramétrique et intrinsèque à l'ensemble E des objets, conformément aux h.a.l. N_1, N_2 et N_3. Chacune de ces h.a.l. associera à l'ensemble $\mathcal{A}$ des attributs de description, un ensemble $\mathcal{A}'$ d'attributs aléatoires indépendants et par conséquent, la famille des indices aléatoires d'association :

$$\{Q_s(a',b')/\{a',b'\} \in P_2(\mathcal{A}')\}, \qquad (17)$$

où

$$Q_s(a',b') = Q(a',b') \Big/ \left[\frac{1}{\binom{m}{2}} \sum \{Q^2(a',b')/\{a',b'\} \in P_2(\mathcal{A}')\}\right]^{1/2} \qquad (18)$$

Cette justification s'effectuera ensuite par rapport au point de vue inférentiel introduit au commencement. Ceci nous conduira à de nouvelles formes de réduction globale.

L'étude inférentielle a été introduite et extensivement développée dans les travaux de L. Goodman et W. Kruskal [GOODMAN et KRUSKAL(1963),(1972)] relativement à des indices que ces auteurs proposent et particulièrement à leur coefficient γ entre deux variables qualitatives ordinales. Toutefois et curieusement, ils ne considèrent pas le cas -que nous verrons assez riche- de la comparaison de deux variables logiques d'incidence (i.e. attributs descriptifs) pour lesquelles il y avait un coefficient aussi bien établi que celui de K. Pearson. D'autre part, les bases formelle et statistique de l'expression des indices mentionnés -et on s'en rend compte lorsqu'il s'agit de les définir au niveau de la population parente dont provient l'échantillon

sur lequel les variables ont été effectivement observées- ne sont pas clairement établies. Nous y reviendrons. Enfin, les auteurs précités n'étudient que la comparaison de deux variables qualitatives seulement ; alors que notre problème est plus général, voire de nature statistique différente et concerne la comparaison deux à deux d'un ensemble de variables observées. C'est naturellement que nous avons été amenés à ce problème, en raison de l'étude de la stabilité -pour n=card(E) croissant- de la classification hiérarchique de la vraisemblance du lien portée sur un ensemble de variables de description d'un champ donné. Des résultats expérimentaux intéressants ont été obtenus dans le cadre d'un D.E.A. [BLANCARD(1976)].

Dans ces conditions, le plan de notre présentation sera le suivant :

II. ASSOCIATION ENTRE DEUX ATTRIBUTS PAR RAPPORT A UN ECHANTILLON DE TAILLE CROISSANTE.

II.1. INTRODUCTION.

Désignons par $\{\pi(a\wedge b),\pi(a\wedge\bar{b}),\pi(\bar{a}\wedge b),\pi(\bar{a}\wedge\bar{b})\}$ la distribution jointe du couple (a,b) d'attributs sur la population totale et parente $\mathcal{P}$ qui définit un ensemble de cardinal N en général "très grand" mais fini. Ainsi, $\pi(a\wedge b)=\mathcal{N}(a\wedge b)/\mathcal{N}$ où $\mathcal{N}(a\wedge b)$ est le nombre d'individus de $\mathcal{P}$ possédant les deux attributs a et b. De même, on définit $\pi(a\wedge\bar{b})=\mathcal{N}(a\wedge\bar{b})/\mathcal{N}$, $\pi(\bar{a}\wedge b)=\mathcal{N}(\bar{a}\wedge b)/\mathcal{N}$ et $\pi(\bar{a}\wedge\bar{b})=\mathcal{N}(\bar{a}\wedge\bar{b})/\mathcal{N}$. Enfin, $\pi(a)=\mathcal{N}(a)/\mathcal{N}$, $\pi(\bar{a})=\mathcal{N}(\bar{a})/\mathcal{N}$, $\pi(b)=\mathcal{N}(b)/\mathcal{N}$ et $\pi(\bar{b})=\mathcal{N}(\bar{b})/\mathcal{N}$, avec des notations que l'on comprend.

Nous supposons que E est extrait de $\mathcal{P}$ selon un modèle multinomial à quatre catégories : $a\wedge b$, $a\wedge\bar{b}$, $\bar{a}\wedge b$ et $\bar{a}\wedge\bar{b}$, représentées au niveau de $\mathcal{P}$ avec les proportions théoriques $\pi(a\wedge b)$, $\pi(a\wedge\bar{b})$, $\pi(\bar{a}\wedge b)$ et $\pi(\bar{a}\wedge\bar{b})$. $p(a\wedge b)$, $p(a\wedge\bar{b})$, $p(\bar{a}\wedge b)$ et $p(\bar{a}\wedge\bar{b})$ sont les proportions observées au niveau de l'échantillon E supposé de taille n. $F(a\wedge b)$, $F(a\wedge\bar{b})$, $F(\bar{a}\wedge b)$ et $F(\bar{a}\wedge\bar{b})$ sont les v.a. respectivement associées à $p(a\wedge b)$, $p(a\wedge\bar{b})$, $p(\bar{a}\wedge b)$ et $p(\bar{a}\wedge\bar{b})$; ainsi ${}^t(nF(a\wedge b),nF(a\wedge\bar{b}),nF(\bar{a}\wedge b),nF(\bar{a}\wedge\bar{b}))$ -où ^{t} désigne transposé- est un vecteur multinomial d'espérance mathématique ${}^t(n\pi(a\wedge b),n\pi(a\wedge\bar{b}),n\pi(\bar{a}\wedge b),n\pi(\bar{a}\wedge\bar{b}))$ et dont la matrice des variances est nW où

$$W=\left[\begin{array}{c|c|c|c}
\pi(a\wedge b)[1-\pi(a\wedge b)] & -\pi(a\wedge b)\pi(a\wedge\bar{b}) & -\pi(a\wedge b)\pi(\bar{a}\wedge b) & -\pi(a\wedge b)\pi(\bar{a}\wedge\bar{b}) \\
-\pi(a\wedge\bar{b})\pi(a\wedge b) & -\pi(a\wedge\bar{b})[1-\pi(a\wedge\bar{b})] & -\pi(a\wedge\bar{b})\pi(\bar{a}\wedge b) & -\pi(a\wedge\bar{b})\pi(\bar{a}\wedge\bar{b}) \\
-\pi(\bar{a}\wedge b)\pi(a\wedge b) & -\pi(\bar{a}\wedge b)\pi(a\wedge\bar{b}) & \pi(\bar{a}\wedge b)[1-\pi(\bar{a}\wedge b)] & -\pi(\bar{a}\wedge b)\pi(\bar{a}\wedge\bar{b}) \\
-\pi(\bar{a}\wedge\bar{b})\pi(a\wedge b) & -\pi(\bar{a}\wedge\bar{b})\pi(a\wedge\bar{b}) & -\pi(\bar{a}\wedge\bar{b})\pi(\bar{a}\wedge b) & \pi(\bar{a}\wedge\bar{b})[1-\pi(\bar{a}\wedge\bar{b})]
\end{array}\right] \quad (1)$$

Dans ces conditions, pour n "assez grand" et avec une excellente approximation, le vecteur aléatoire des fréquences relatives ${}^t(F(a\wedge b),F(a\wedge\bar{b}),F(\bar{a}\wedge b),F(\bar{a}\wedge\bar{b}))$ suit une loi 4-normale dont les paramètres (vecteur moyen et matrice des variances) se déduisent immédiatement de ci-dessus.

Nous avons déjà exprimé les indices ${}_1r(a,b)$, ${}_2r(a,b)$ et ${}_3r(a,b)$ (cf.(2), (5) et (8) du paragraphe I ci-dessus).

Nous pouvons -exactement à partir de la même démarche, mais au niveau de la population entière - déterminer les indices correspondants ${}_1\rho(a,b)$, ${}_2\rho(a,b)$ et ${}_3\rho(a,b)$ qui ont, respectivement, les mêmes expressions que ${}_1r(a,b)$, ${}_2r(a,b)$ et ${}_3r(a,b)$, mais où, les proportions p observées au niveau de l'échantillon E, sont remplacées par les proportions théoriques π définis au niveau de la population totale $\mathcal{P}$.

Il nous reste à introduire les v.a. ${}_1R(a,b)$ ${}_2R(a,b)$ et ${}_3R(a,b)$ dont les expressions sont respectivement identiques à celles de ${}_1r(a,b)$, ${}_2r(a,b)$ et ${}_3r(a,b)$, à celà près qu'on remplace les fréquences relatives observées $p(a\wedge b)$, $p(a\wedge\bar{b})$, $p(\bar{a}\wedge b)$ et $p(\bar{a}\wedge\bar{b})$ par, respectivement, celles aléatoires : $F(a\wedge b)$, $F(a\wedge\bar{b})$, $F(\bar{a}\wedge b)$ et $F(\bar{a}\wedge\bar{b})$.

Nous allons maintenant étudier la distribution asymptotique de ${}_\varepsilon R(a,b)$. En raison de la remarque qui suit (10) du paragraphe I, nous nous limitons à $\varepsilon=1$ et $\varepsilon=3$.

II.2. DISTRIBUTIONS ASYMPTOTIQUES DE ${}_1R(a,b)$ et de ${}_3R(a,b)$; étude de la variance.

Ces distributions s'obtiennent directement à partir de l'application d'un théorème général (cf. S.WILKS(1962) p.260) valable pour n'importe quelle dimension finie k, que nous allons exprimer pour les besoins de notre cause dans le cas où k=4 et ce, en utilisant les notations les plus suggestives.

THEOREME. Soit $\{(s_i,u_i,v_i,t_i)/1\leqslant i\leqslant n\}$ un échantillon de taille n provenant d'une distribution de dimension 4 dont le vecteur moyen est ${}^t(\pi_{11},\pi_{10},\pi_{01},\pi_{00})$ et dont la matrice définie positive des variances est $\mathcal{W}$. Soit $\gamma(s,u,v,t)$ une fonction qui possède des dérivées premières $\frac{\partial\gamma}{\partial s}$, $\frac{\partial\gamma}{\partial u}$, $\frac{\partial\gamma}{\partial v}$ et $\frac{\partial\gamma}{\partial t}$ dans le voisinage du point $(\pi_{11},\pi_{10},\pi_{01},\pi_{11})$ et posons $\gamma_s^o,\gamma_u^o,\gamma_v^o$ et γ_t^o pour les valeurs respectives de ces dérivées partielles en ce point. Si l'une au moins de ces valeurs est différente de zéro, alors la v.a. $\gamma(\bar{S},\bar{U},\bar{V},\bar{T})$ -où $\bar{S},\bar{U},\bar{V}$, et $\bar{T}$ représentent respectivement les moyennes d'échantillonnage de la suite des quatre variables- a une distribution asymptotique normale de moyenne $\gamma(\pi_{11},\pi_{10},\pi_{01},\pi_{00})$ et de variance $\frac{1}{n}({}^t\gamma^o W\gamma^o)$ où nous notons ${}^t\gamma^o=(\gamma_s^o,\gamma_u^o,\gamma_v^o,\gamma_t^o)$.

Dans notre problème (s_i,u_i,v_i,t_i) qui représente le i-ème tirage dans une urne multinomiale est un vecteur logique formé de trois 0 et d'un seul 1. $\pi_{11},\pi_{10},\pi_{01}$ et π_{11} sont les proportions définies au niveau de la population $\mathcal{P}$, précédemment, respectivement notées, $\pi(a\wedge b),\pi(a\wedge\bar{b}),\pi(\bar{a}\wedge b)$ et $\pi(\bar{a}\wedge\bar{b})$. Quant à la fonction γ, elle sera définie par l'expression de ${}_1\rho$ ou celle de ${}_3\rho$; de sorte que ${}_1R(a,b)$ (resp. ${}_1R(a,b)$) correspondra à une v.a. de même nature que $\gamma(\bar{S},\bar{U},\bar{V},\bar{T})$.

PROPRIETE. La distribution asymptotique de ${}_{\varepsilon}R(a,b)$ (ε=1 ou 3) est normale de moyenne ${}_{\varepsilon}\rho(a,b)$ et de variance $\frac{1}{n}({}^{t}{}_{\varepsilon}\rho^{o} W {}_{\varepsilon}\rho^{o})$ où W est la matrice définie par la formule (22) ci-dessus et où ${}^{t}{}_{\varepsilon}\rho^{o}=({}_{\varepsilon}\rho^{o}_{11}, {}_{\varepsilon}\rho^{o}_{10}, {}_{\varepsilon}\rho^{o}_{01}, {}_{\varepsilon}\rho^{o}_{00})$, avec ${}_{\varepsilon}\rho^{o}_{11} = \frac{\partial_{\varepsilon}\rho}{\partial\pi_{11}}$, ${}_{\varepsilon}\rho^{o}_{10} = \frac{\partial_{\varepsilon}\rho}{\partial\pi_{10}}$, ${}_{\varepsilon}\rho^{o}_{01} = \frac{\partial_{\varepsilon}\rho}{\partial\pi_{01}}$ et ${}_{\varepsilon}\rho^{o}_{00} = \frac{\partial_{\varepsilon}\rho}{\partial\pi_{00}}$ pris au point $(\pi_{11},\pi_{10},\pi_{01},\pi_{00})$.

On démontrera sans peine que, dans le cas non dégénéré où chacune des proportions théoriques $\pi_{11},\pi_{10},\pi_{01}$ et π_{00} est différente de zéro, l'un au moins des quatre nombres ${}_{\varepsilon}\rho^{o}_{11}$, ${}_{\varepsilon}\rho^{o}_{10}$, ${}_{\varepsilon}\rho^{o}_{01}$ et ${}_{\varepsilon}\rho^{o}_{00}$ est différent de zéro et ce, aussi bien pour ε=1 que pour ε=3.

COROLLAIRE. La variance de ${}_{\varepsilon}R(a,b)$ est, au facteur (1/n) près, la variance d'une v.a. discrète ${}_{\varepsilon}R^{*}$ pouvant prendre l'une des valeurs $\rho^{o}_{11},\rho^{o}_{10},\rho^{o}_{01}$ et ρ^{o}_{00} avec respectivement, les probabilités $\pi_{11},\pi_{10},\pi_{01}$ et π_{00}.

En effet, considérons le vecteur aléatoire (S_i,U_i,V_i,T_i) -associé à l'extraction du i-ème individu de l'urne 4-nomiale- dont la suite des valeurs possibles est $((1,0,0,0),(0,1,0,0),(0,0,1,0),(0,0,0,1))$ où la suite des probabilités respectivement affectées est $(\pi_{11},\pi_{10},\pi_{01},\pi_{00})$.

Le vecteur aléatoire ${}^{t}(S_i,U_i,V_i,T_i)$ a précisément W pour matrice des variances. On peut aisément se rendre compte que la variance de la variable aléatoire

$$\rho^{o}_{11} S_i + \rho^{o}_{10} U_i + \rho^{o}_{01} V_i + \rho^{o}_{11} T_i, \qquad (2)$$

peut se mettre sous la forme : ${}^{t}\rho^{o} W \rho^{o}$

On retrouve ainsi et de façon sensiblement plus synthétique le résultat de [GOODMAN & KRUSKAL(1972)].

II.2.1. Moyenne et Variance de ${}_{1}R^{*}$.

L'objet de ce paragraphe est de toucher de plus près les valeurs de la moyenne et de la variance de ${}_{1}R^{*}$.

Pour alléger les écritures, nous allons ici noter par γ l'indice ${}_{1}\rho$ et par, respectivement, σ,δ,η et θ les proportions $\pi_{11},\pi_{10},\pi_{01}$ et π_{00} ; enfin, nous désignerons par ϕ l'expression $\pi(a)\pi(\bar{a})\pi(b)\pi(\bar{b}) = (\pi_{11}+\pi_{10})(\pi_{01}+\pi_{00})(\pi_{11}+\pi_{01})(\pi_{10}+\pi_{00})$

$= (\sigma+\delta)(\eta+\theta)(\sigma+\eta)(\delta+\theta) = \phi$.

THEOREME 1. L'espérance mathématique de ${}_1R^*$ est nulle.

On peut mettre $\partial\gamma/\partial\sigma$, $\partial\gamma/\partial\delta$, $\partial\gamma/\partial\eta$ et $\partial\gamma/\partial\theta$ sous la forme suivante :

$$\left.\begin{aligned}
&\frac{\partial\gamma}{\partial\sigma}\ \frac{1}{2\sqrt{\phi}}\left\{-\frac{[\pi(a)\pi(\bar b)+\pi(\bar a)\pi(b)]}{\pi(a)\pi(b)}\times\pi(a\wedge b)+[\pi(\bar a)+\pi(\bar b)]\right\}\\
&\frac{\partial\gamma}{\partial\theta}\ \frac{1}{2\sqrt{\phi}}\left\{-\frac{[\pi(\bar a)\pi(b)+\pi(a)\pi(\bar b)]}{\pi(\bar a)\pi(\bar b)}\times\pi(\bar a\wedge\bar b)+[\pi(a)+\pi(b)]\right\}\\
&\frac{\partial\gamma}{\partial\delta}\ \frac{1}{2\sqrt{\phi}}\left\{\frac{[\pi(a)\pi(b)+\pi(\bar a)\pi(\bar b)]}{\pi(a)\pi(\bar b)}\times\pi(a\wedge\bar b)+[\pi(\bar a)+\pi(b)]\right\}\\
&\frac{\partial\gamma}{\partial\eta}\ \frac{1}{2\sqrt{\phi}}\left\{\frac{[\pi(\bar a)\pi(\bar b)+\pi(a)\pi(b)]}{\pi(\bar a)\pi(\bar b)}\times\pi(\bar a\wedge b)+[\pi(a)+\pi(\bar b)]\right\}
\end{aligned}\right\}\qquad(3)$$

Dans ces conditions, il y a lieu d'établir que la somme suivante est nulle :

$$\left.\begin{aligned}
&\left\{-\frac{[\pi(a)\pi(\bar b)+\pi(\bar a)\pi(b)]}{\pi(a)\pi(b)}\times\pi^2(a\wedge b)+[\pi(a)+\pi(\bar b)]\pi(a\wedge b)\right\}\\
&\overset{+}{\left\{-\frac{|\pi(\bar a)\pi(b)+\pi(a)\pi(\bar b)|}{\pi(\bar a)\pi(\bar b)}\times\pi^2(\bar a\wedge\bar b)+[\pi(a)+\pi(b)]\pi(\bar a\wedge\bar b)|\right\}}\\
&\overset{+}{\left\{\frac{[\pi(a)\pi(b)+\pi(\bar a)\pi(\bar b)]}{\pi(a)\pi(\bar b)}\times\pi^2(a\wedge\bar b)-[\pi(\bar a)+\pi(b)]\pi(a\wedge\bar b)|\right\}}\\
&\overset{+}{\left\{\frac{[\pi(\bar a)\pi(\bar b)+\pi(a)\pi(b)]}{\pi(\bar a)\pi(b)}\times\pi^2(\bar a\wedge b)-[\pi(a)+\pi(\bar b)]\pi(\bar a\wedge b)|\right\}}
\end{aligned}\right\}\qquad(4)$$

Pour ne pas alourdir ce texte, nous laissons le soin au lecteur d'effectuer le détail des calculs en nous contentant de signaler leur organisation générale. Commençons par noter les relations suivantes permettant d'exprimer tous les éléments par rapport à $\pi(a)$, $\pi(b)$ et $\pi(a\wedge b)$:

$$\left.\begin{aligned}
&\pi(\bar a)=1-\pi(a),\quad \pi(\bar b)=1-\pi(b)\\
&\pi(a\wedge\bar b)=\pi(a)-\pi(a\ b),\quad \pi(\bar a\wedge b)=\pi(b)-\pi(a\wedge b)\\
\text{et}\quad&\pi(\bar a\wedge\bar b)=\pi(a\wedge b)-\pi(a)-\pi(b)+1
\end{aligned}\right\}\qquad(5)$$

Mais, au préalable, le calcul va s'organiser en établissant les contributions respectives de la première puis de la deuxième colonne de la disposition (4). La contribution de la première colonne est, au facteur $1/\pi(a)\pi(\bar a)\pi(b)\pi(\bar b)$ près, égale à

$$\begin{aligned}
&[\pi(a)\pi(b)+\pi(\bar a)\pi(\bar b)]\{\pi(\bar a)\pi(b)\pi^2(a\wedge\bar b)+\pi(a)\pi(\bar b)\pi^2(\bar a\wedge b)\}\\
-&[\pi(a)\pi(\bar b)+\pi(\bar a)\pi(b)]\{\pi(\bar a)\pi(\bar b)\pi^2(a\wedge b)+\pi(a)\pi(b)\pi^2(\bar a\wedge\bar b)\}\qquad(6)
\end{aligned}$$

Celle de la deuxième colonne de la disposition (4), est égale à

$$\pi(a\wedge b)\left[\pi(\bar{a})+\pi(\bar{b})\right]+\pi(\bar{a}\wedge\bar{b})\left[\pi(a)+\pi(b)\right]-\pi(a\wedge\bar{b})\left[\pi(\bar{a})+\pi(b)\right]-\pi(\bar{a}\wedge b)\left[\pi(a)+\pi(\bar{b})\right]$$
$$=4\left[\pi(a\wedge b)-\pi(a)\pi(b)\right], \quad (7)$$

en tenant compte des relations (5).

Toujours en utilisant les relations (5) , on développe l'intérieur des accolades de l'expression (6) par rapport à $\pi(a\wedge b)$ et en simplifiant après un minutieux calcul, on obtient pour l'expression (6)

$$-4\pi(a)\pi(\bar{a})\pi(b)\pi(\bar{b})\left[\pi(a\wedge b)-\pi(a)\pi(b)\right], \quad (8)$$

ce qui achève d'établir le résultat annoncé.

Avant d'énoncer le deuxième théorème, dont le but est de fournir l'expression la plus synthétique que nous ayons trouvé de la variance de la v.a. ${}_1R(a,b)$, précisons quelques notations intermédiaires. Soit (α,β) un couple variable d'attributs dont l'ensemble des valeurs est $\{(a,b),(a,\bar{b}),(\bar{a},b),(\bar{a},\bar{b})\}=\{a,\bar{a}\}x\{b,\bar{b}\}$, relativement à (α,β), on posera :

$d(\alpha,\beta)=\pi(\alpha\wedge\beta)/\pi(\alpha)\pi(\beta)$: densité en (,),

$H(\alpha,\beta)=\pi(\alpha\wedge\beta)d(\alpha,\beta)d(\bar{\alpha},\bar{\beta})$, $K(\alpha,\beta)=\left[\pi(\alpha)+\pi(\beta)\right]^2\pi(\bar{\alpha})\pi(\bar{\beta})d(\alpha,\beta)$

$\Phi=\sum\{H(\alpha,\beta)/(\alpha,\beta)\in AxB\}$ où $A=\{a,\bar{a}\}$ et $B=\{b,\bar{b}\}$,

et

$\psi=\sum\{K(\alpha,\beta)/(\alpha,\beta)\in AxB\}$. (9)

Avec ces notations, on a

THEOREME 2. La variance de ${}_1R(a,b)$ est égale, au facteur $(1/n)$ près, à

$\Phi+\frac{\gamma^2}{4\phi}\{\psi-4\left[\pi(a)\pi(\bar{a})+\pi(b)\pi(\bar{b})\right]\}$, (10) où rappelons-le, γ représente l'indice ${}_1\rho(a,b)$ et ϕ, le produit $\pi(a)\pi(\bar{a})\pi(b)\pi(\bar{b})$.

Compte-tenu du théorème 1 ci-dessus, cette variance se réduit à

$$\pi(a\wedge b)\left(\frac{\partial\gamma}{\partial\sigma}\right)^2+\pi(a\wedge\bar{b})\left(\frac{\partial\gamma}{\partial\delta}\right)^2+\pi(\bar{a}\wedge b)\left(\frac{\partial\gamma}{\partial\eta}\right)^2+\pi(\bar{a}\wedge\bar{b})\left(\frac{\partial\gamma}{\partial\theta}\right)^2. \quad (11)$$

En nous reportant aux expressions (3) ci-dessus, on a :

$$\left(\frac{\partial\gamma}{\partial\gamma}\right)^2 = \frac{\pi^2(\bar{a}\wedge\bar{b})}{\phi} + \gamma^2 \times \frac{[\pi(a)+\pi(b)]^2}{4\pi^2(a)\pi^2(b)} - \frac{\gamma}{\sqrt{\phi}} \times \frac{[\pi(a)+\pi(b)]}{\pi(a)\pi(b)} \times \pi(\bar{a}\wedge\bar{b}) \qquad \pi(a\wedge b)$$

$$\left(\frac{\partial\gamma}{\partial\theta}\right)^2 = \frac{\pi^2(a\wedge b)}{\phi} + \gamma^2 \times \frac{[\pi(\bar{a})+\pi(\bar{b})]^2}{4\pi^2(\bar{a})\pi^2(\bar{b})} - \frac{\gamma}{\sqrt{\phi}} \times \frac{[\pi(\bar{a})+\pi(\bar{b})]}{\pi(\bar{a})\pi(\bar{b})} \times \pi(a\wedge b) \qquad \pi(\bar{a}\wedge\bar{b})$$

$$\left(\frac{\partial\gamma}{\partial\delta}\right)^2 = \frac{\pi^2(\bar{a}\wedge b)}{\phi} + \gamma^2 \times \frac{[\pi(a)+\pi(\bar{b})]^2}{4\pi^2(a)\pi^2(\bar{b})} - \frac{\gamma}{\sqrt{\phi}} \times \frac{[\pi(a)+\pi(\bar{b})]}{\pi(a)\pi(\bar{b})} \times \pi(\bar{a}\wedge b) \qquad \pi(a\wedge\bar{b})$$

$$\left(\frac{\partial\gamma}{\partial\theta}\right)^2 = \frac{\pi^2(a\wedge\bar{b})}{\phi} + \gamma^2 \times \frac{[\pi(\bar{a})+\pi(b)]}{4\pi^2(\bar{a})\pi^2(b)} - \frac{\gamma}{\sqrt{\phi}} \times \frac{[\pi(\bar{a})+\pi(b)]}{\pi(\bar{a})\pi(b)} \times \pi(a\wedge\bar{b}) \qquad \pi(\bar{a}\wedge b)$$

① ② ③ (12)

Relativement à la disposition (12) des calculs, on voit clairement que la contribution de la colonne ① à l'expression (11) est égale à Φ, que celle de la colonne ② est égale à $(\gamma^2\psi/4\phi)$. Enfin, la contribution de la colonne 3 nécessite calcul et simplification la ramenant à la forme $(-\gamma^2[\pi(a)\pi(\bar{a})+\pi(b)\pi(\bar{b})]/\phi)$.

CORROLAIRE. Si les deux attributs a et b sont indépendants, la variance de ${}_1R(a,b)$ est alors égale à $(1/n)$.

Ce cas est caractérisé par $d(a,b)=d(a,\bar{b})=d(\bar{a},b)=d(\bar{a},\bar{b})=1$ où l'une des relations $d(\alpha,\beta)=1$, est équivalente aux trois autres de même forme, ce qui implique clairement que $\Phi=1$. D'autre part, dans ce cas $\gamma=0$. D'où le résultat.

On remarquera que dans le cas -qui n'a pas de sens réel- où l'attribut b est identique à l'attribut a, la variance de ${}_1R(a,b)$ est nécessairement nulle puisque ${}_1R(a,b)$ est identiquement égal à 1. Pour des raisons de cohérence, on le vérifie d'ailleurs aisément sur l'expression (10) ci-dessus où alors $\Phi=1/\pi(a)\pi(a)$, $\gamma^2=1$, $\phi=\pi^2(a)\pi^2(a)$ et $\psi=4\pi(a)\pi(a)$.

Pour terminer, nous allons illustrer la valeur de var.$({}_1R(a,b))$ donné par l'expression (10), dans quelques situations particulières :

$\pi(a\wedge b)$	$\pi(a\wedge\bar{b})$	$\pi(\bar{a}\wedge b)$	$\pi(\bar{a}\wedge\bar{b})$	$\pi(a)$	$\pi(\bar{a})$	$\pi(b)$	$\pi(\bar{b})$	var$[{}_1R(a,b)]$
0,10	0,40	0,40	0,10	0,50	0,50	0,50	0,50	0,64/n
0,20	0,30	0,30	0,20	0,50	0,50	0,50	0,50	0,96/n
0,125	0,125	0,125	0,625	0,25	0,75	0,25	0,75	1,136/n
0,125	0	0,125	0,75	0,125	0,875	0,25	0,75	0,57/n

On se rend compte, mais il y a lieu de le préciser par une tabulation plus importante, que comme on peut s'y attendre intuitivement, la variance faiblit dans le cas de forte liaison (voir ligne 1 du tableau ci-dessus et surtout ligne 4 qui correspond à une inclusion totale $E(a) \subset E(b)$ et semble avoisiner la valeur $(1/n)$ autour de l'indépendance. On peut certes donner une valeur approximative de (10) en remplaçant les proportions théoriques par celles estimées au niveau de l'échantillon E. Mais il semble que dans la quasi-totalité des cas, cette variance soit comprise entre $0,5/n$ et $1,5/n$.

II.2.2. Moyenne et Variance de ${}_3R(a,b)$

Nous prenons ici la notation λ pour le coefficient ${}_3\rho$.

THEOREME 3. La moyenne et la variance de l'estimateur ${}_3R(a,b)$ sont respectivement égales à

$$\lambda(a,b) \text{ et } \frac{1}{\eta\pi(a)\pi(b)} \{\pi(a\wedge b)\pi(\bar{a}\wedge\bar{b})\left[\pi(a\wedge b)+\pi(\bar{a}\wedge\bar{b})\right]+\pi(a\wedge\bar{b})\pi(\bar{a}\wedge b)\left[\pi(a\wedge\bar{b})+\pi(\bar{a}\wedge b)\right] - \frac{1}{4}\lambda^2(a,b)\left[\left[\pi(a\wedge\bar{b})+\pi(\bar{a}\wedge b)\right]+2\left[\pi(a)+\pi(b)+2\pi(a)\pi(b)\right]\right]\} \qquad (13)$$

Conformément au paragraphe précédent, nous allons calculer les dérivées partielles $(\partial\lambda/\partial\sigma)$, $(\partial\lambda/\partial\theta)$, $(\partial\lambda/\partial\delta)$ et $(\partial\lambda/\partial\eta)$ prises au point $(\pi(a\wedge b), \pi(a\wedge\bar{b}), \pi(\bar{a}\wedge b), \pi(a\wedge\bar{b}))$:

$$\left.\begin{array}{l|l} \dfrac{\partial\lambda}{\partial\sigma} = \dfrac{\pi(\bar{a}\wedge\bar{b})}{\sqrt{\pi(a)\pi(b)}} - \lambda\dfrac{\left[\pi(a)+\pi(b)\right]}{2\pi(a)\pi(b)} & \dfrac{\partial\lambda}{\partial\delta} = \dfrac{-\pi(\bar{a}\wedge b)}{\sqrt{\pi(a)\pi(b)}} - \lambda\times\dfrac{\pi(b)}{2\pi(a)\pi(b)} \\ \dfrac{\partial\lambda}{\partial\theta} = \dfrac{\pi(a\wedge b)}{\sqrt{\pi(a)\pi(b)}} & \dfrac{\partial\lambda}{\partial\eta} \quad \dfrac{-\pi(a\wedge\bar{b})}{\sqrt{\pi(a)\pi(b)}} - \lambda * \dfrac{\pi(a)}{2\pi(a)\pi(b)} \end{array}\right\} \qquad (14)$$

On obtient pour la moyenne : $\left[\pi(a\wedge b)(\partial\lambda/\partial\sigma)+\pi(a\wedge\bar{b})(\partial\lambda/\partial\delta)+\pi(\bar{a}\wedge b)(\partial\lambda/\partial\eta)+\pi(\bar{a}\wedge\bar{b})(\partial\lambda/\partial\theta)\right]$, exactement la valeur λ.

Il reste maintenant à déterminer la moyenne des carrés :

$$\pi(a\wedge b)(\partial\lambda/\partial\sigma)^2+\pi(a\wedge\bar{b})(\partial\lambda/\partial\delta)^2+\pi(\bar{a}\wedge b)(\partial\lambda/\partial\eta)^2+\pi(\bar{a}\wedge\bar{b})(\partial\lambda/\partial\theta)^2 \qquad (15)$$

$$\left.\begin{array}{ll} \left(\dfrac{\partial\lambda}{\partial\sigma}\right)^2 = \dfrac{1}{\pi(a)\pi(b)}\{\pi^2(\bar{a}\wedge\bar{b}) - \dfrac{\left[\pi(a)+\pi(b)\right]\pi(\bar{a}\wedge\bar{b})}{\sqrt{\pi(a)\pi(b)}}\times\lambda + \dfrac{\left[\pi(a)+\pi(b)\right]^2}{4\,\pi(a)\pi(b)}\lambda^2\} & \pi(a\wedge b) \\ \left(\dfrac{\partial\lambda}{\partial\delta}\right)^2 = \dfrac{1}{\pi(a)\pi(b)}\{\pi^2(\bar{a}\wedge b) + \dfrac{\pi(b)\pi(\bar{a}\wedge b)}{\sqrt{\pi(a)\pi(b)}}\times\lambda + \dfrac{\left[\pi(b)\right]^2}{4\,\pi(a)\pi(b)}\lambda^2\} & \pi(a\wedge\bar{b}) \\ \left(\dfrac{\partial\lambda}{\partial\eta}\right)^2 = \dfrac{1}{\pi(a)\pi(b)}\{\pi^2(a\wedge\bar{b}) + \dfrac{\pi(a)\pi(a\wedge\bar{b})}{\sqrt{\pi(a)\pi(b)}}\times\lambda + \dfrac{\left[\pi(a)\right]^2}{4\,\pi(a)\pi(b)}\} & \pi(\bar{a}\wedge b) \\ \left(\dfrac{\partial\lambda}{\partial\theta}\right) = \dfrac{\pi^2(a\wedge b)}{\pi(a)\pi(b)} & \pi(\bar{a}\wedge\bar{b}) \end{array}\right\} \qquad (16)$$

① ② ③

La contribution de la colonne ① est

$\frac{1}{\pi(a)\pi(b)} \{\pi(a\wedge b)\pi(\bar{a}\wedge\bar{b})\left[\pi(a\wedge b)+\pi(\bar{a}\wedge\bar{b})\right] + \pi(a\wedge\bar{b})\pi(\bar{a}\wedge b)\left[\pi(a\wedge\bar{b})+\pi(\bar{a}\wedge b)\right]\}$, celle de la colonne ② peut se réduire à

$-\frac{\left[\pi(a)+\pi(b)\right]}{\pi(a)\pi(b)}\ \lambda^2$, enfin celle de la colonne ③ vaut

$$\frac{\left[\pi(a)+\pi(b)+2\pi(a\wedge b)\right]}{4\ \pi(a)\pi(b)} \times \lambda^2$$

Ainsi, la somme des contributions ② et ③ peut se mettre sous la forme

$$\frac{-\lambda^2}{4\ \pi(a)\pi(b)} \{\left[\pi(a\wedge\bar{b})+\pi(\bar{a}\wedge b)\right]+2\left[\pi(a)+\pi(b)\right]\}$$

D'où, la variance des nombres de (14) relativement à $\{\pi(\alpha,\beta)/(\alpha,\beta)\in AxB\}$:

$$\frac{1}{\pi(a)\pi(b)} \{\pi(a\wedge b)\pi(\bar{a}\wedge\bar{b})\left[\pi(a\wedge b)+\pi(\bar{a}\wedge\bar{b})\right] + \pi(a\wedge\bar{b})\pi(\bar{a}\wedge b)\left[\pi(a\wedge\bar{b})+\pi(\bar{a}\wedge b)\right] - \frac{\lambda^2}{4}\left[\left[\pi(a\wedge\bar{b})+\pi(\bar{a}\wedge b)\right]+2\left[\pi(a)+\pi(b)+2\pi(a)\pi(b)\right]\right]\}, \quad (17)$$

ce qui donne l'expression annoncée de var $\left[{}_3R(a,b)\right]$.

Nous allons nous contenter d'illustrer numériquement une fois l'expression (17).

Considérons l'exemple suivant : $\pi(a\wedge b)=0,4$, $\pi(a\wedge\bar{b})=0,1$, $\pi(\bar{a}\wedge b)=0,1$ et $\pi(\bar{a}\wedge\bar{b})=0,4$. On obtient la valeur 0,232 pour l'expression (17).

III. *COMPARAISON DE DEUX PAIRES D'ATTRIBUTS ; Estimation et précision calcul en analyse des données.*

III.1. INTERVALLES DE CONFIANCE ET COMPARAISON ORDINALE DES ASSOCIATIONS.

Désignons par C^2/n et par L^2/n les variances respectives de ${}_1R(a,b)$ et de ${}_3R(a,b)$ (cf. expressions (10) et (13), §II).

Compte tenu du caractère asymptotiquement normal de la distribution de ${}_\varepsilon R(a,b)$ (cf. "Propriété" ci-dessus), on a, à partir de l'observation de ${}_\varepsilon r(a,b)$ ($\varepsilon=1$ ou $\varepsilon=3$), les intervalles de confiance symétriques au seuil $(1-\alpha)$ pour ${}_1\rho$ et ${}_3\rho$:

$$\left[{}_1r(a,b)-\sqrt{C^2/n}\ \ G^{-1}(1-\tfrac{\alpha}{2}),\ {}_1r(a,b)+\sqrt{C^2/n}\ \ G^{-1}(1-\tfrac{\alpha}{2})\right] \quad (1)$$

et

$$\left[{}_3r(a,b)-\sqrt{L^2/n}\ \ G^{-1}(1-\tfrac{\alpha}{2}),\ {}_3r(a,b)+\sqrt{L^2/n}\ \ G^{-1}(1-\tfrac{\alpha}{2})\right], \quad (2)$$

où G est la f.r. de la loi normale centrée réduite.

En considérant un seuil de confiance de l'ordre de 0,99, les intervalles de confiance (45) et (46) deviennent, lorsque a et b sont (au niveau de la population parente $\mathcal{P}$) indépendants, voire même "faiblement"(cf. dernier tableau du paragraphe II ci-dessus) :

$$\left[{}_1r(a,b)-2,5/\sqrt{n},\ {}_1r(a,b)+2,5/\sqrt{n}\right] \qquad (3)$$

$$\left[{}_3r(a,b)-2,5\sqrt{\pi(\bar{a})\pi(\bar{b})}/\sqrt{n},\ {}_3r(a,b)+2,5\sqrt{\pi(\bar{a})\pi(\bar{b})}/\sqrt{n}\right] ; \qquad (4)$$

en effet, on peut voir que $L^2=\pi(\bar{a})\pi(\bar{b})$ dans le cas de l'indépendance.

Ainsi, on se rend compte, qu'en termes d'intervalle de confiance, la valeur de ${}_1r(a,b)$ (resp. ${}_3r(a,b)$) ne peut refléter la valeur exacte de ${}_1\rho(a,b)$ (resp. ${}_3\rho(a,b)$) au delà du premier chiffre significatif et cela même pour $n=10^4$!.

On peut remarquer que si on adopte un seuil de confiance plus petit, par exemple de 0,95 ou même de 0,90, l'ordre de grandeur de l'amplitude de l'intervalle de confiance reste quasiment le même et concerne les premiers chiffres significatifs de l'indice.

C'est et de façon implicite cette dernière circonstance qui rend circonspet puis surprend le statisticien classique devant le succès de l'analyse des données. Toutefois, il faut souligner que le problème n'est pas tant d'estimer la valeur d'un seul indice d'association entre deux variables sur une population -dont d'ailleurs la valeur calculée sur l'échantillon E correspond à l'estimation du maximum de vraisemblance- que de comparer de façon simplement ordinale les corrélations deux à deux d'une famille de variables ou de classes de variables. On sait même qu'il existe des méthodes efficaces de classification et même de représentation euclidienne basées sur la seule "préordonnance" associée à l'indice d'association (i.e. rangement des paires d'éléments de l'ensemble à organiser par ressemblance décroissante). Nous avons toutefois pu tester que la qualité de ces méthodes restait très intimement liée à celle de l'indice de ressemblance ayant permis d'établir la préordonnance.

Dans ces conditions, pour achever d'avoir une idée sur la précision calcul nécessaire, considérons le problème de la comparaison de deux paires d'attributs {a,b} et {c,d} que -pour des raisons de simplicité- nous supposons sans composante commune. Les v.a. asymptotiquement normales ${}_1R(a,b)$ et ${}_1R(c,d)$ sont indépendantes. Imaginons alors une situation numérique où ${}_1\rho(a,b)=0$ et ${}_1\rho(c,d)=0,08$, donc coïncidant (dans le cas de {a,b}) ou voisine (dans le cas de {c,d}) de l'indépendance, mais où tout de même ${}_1\rho(a,b)<{}_1\rho(c,d)$. Dans cette situation, la variance de ${}_1R(a,b)$ est exactement égale à $1/\sqrt{n}$ et celle de ${}_1R(c,d)$ voisine de $1/\sqrt{n}$, nous l'assimilerons également à $1/\sqrt{n}$. Dans ces conditions, la v.a. $\left[{}_1R(c,d)-{}_1R(a,b)\right]$ est normale de moyenne $\left[{}_1\rho(c,d)-{}_1\rho(a,b)\right]$ et de variance $2/n$. De sorte que, pour $n=10^4$:

$$\Pr\{{}_1R(a,b)<{}_1R(c,d)\}=G(8/\sqrt{2})\cong 1, \qquad (5)$$

où G est la f.r. de la loi normale centrée réduite.

Ainsi, c'est avec une quasi-certitude que la relation ${}_1\rho(a,b)<{}_1\rho(c,d)$ se trouvera vérifiée au niveau de l'échantillon E de taille $n=10^4$.

Compte tenu de ce dernier fait, mais en pondérant également par la qualité de la précision des intervalles de confiance, le nombre k de chiffres significatifs après la virgule que nous préconisons pour la mesure des coefficients d'association entre variables, est défini par $10^k \leqslant n < 10^{(k+1)}$. D'où l'aberration des programmes d'analyse des données qui travaillent avec la double précision, alors que l'effectif -d'ailleurs important de l'échantillon- ne permet en fait que deux ou trois chiffres significatifs pour l'estimation des corrélations entre variables...

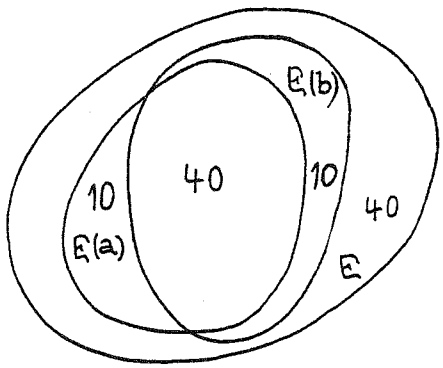
III.2. COMPARAISON ENTRE DEUX ASSOCIATIONS RELATIVES A DEUX ECHANTILLONS DE TAILLES DISTINCTES ; un paradoxe ?

Pour des raisons de simplicité d'interprétation, les deux paires d'attributs {a,b} et {c,d} sont supposées sans composante commune. D'autre part, on considèrera ici l'indice d'expression plus simple $_3\rho(u,v)$ défini au niveau de la population entière , celui correspondant $_3r(u,v)$ défini au niveau de l'échantillon E et enfin celui toujours défini au niveau de E, mais basé sur la vraisemblance de la relation $\Phi\left[\sqrt{n}\ {}_3r(u,v)\right]$ (cf. formule (11) § I.1.), où Φ est la f.r. de la loi normale centrée réduite.

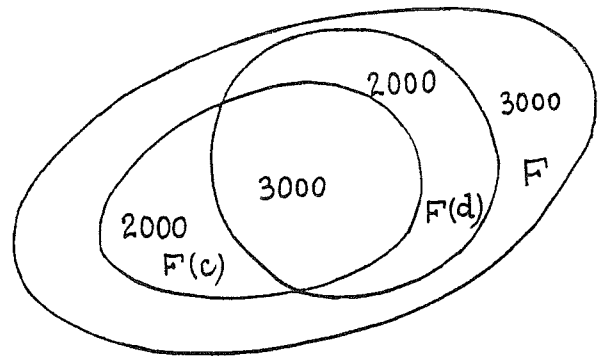
Il s'agit donc de comparer $_3\rho(a,b)$ et $_3\rho(c,d)$ dont les estimations de maximum de vraisemblance sur deux échantillons aléatoires indépendants de tailles respectives n et m, sont notées $_3r_n(a,b)$ et $_3r_m(a,b)$.

Notre but est ici -en nous appuyant sur un exemple- de mettre en évidence un illogisme statistique qui choque le bon sens le plus élémentaire et qui résulte de la philosophie des tests inférentiels (hypothèses définies au niveau de $\mathcal{P}$) d'indépendance. Cet illogisme justifiera a posteriori notre démarche dans l'évaluation des dépendances ou des ressemblances que nous achèverons d'établir au paragraphe suivant.

Imaginons deux échantillons aléatoires indépendants E et F de tailles respectives n=100 et m=10.000. Supposons que l'observation de la paire {a,b} d'attributs sur E, donne lieu à la situation cardinale : $n(a\wedge b)=40$, $n(a\wedge\bar{b})=10$, $n(\bar{a}\wedge b)=10$ et $n(\bar{a}\wedge\bar{b})=40$.



Supposons d'autre part que l'observation de la paire {c,d} d'attributs sur F, donne lieu à la situation cardinale : $m(c\wedge d)=3000$, $m(c\wedge\bar{d})=2000$, $m(\bar{c}\wedge d)=2000$ et $m(\bar{c}\wedge\bar{d})=3000$, schématisée ci- après



La valeur de l'indice $_3r(a,b)$ est :

$$_3r(a,b)=\left[0,4-0,25\right]/\sqrt{0,25}=0,3$$

La valeur de l'indice $_3r(c,d)$ est :

$$_3r(c,d)=\left|0,3-0,25\right|/\sqrt{0,25}=0,1$$

On a $\sqrt{n}\ _3r(a,b)=3$ et $\sqrt{m}\ _3r(c,d)=10$.

Si on considère maintenant dans chacun des deux cas le test de l'hypothèse d'indépendance entre les deux attributs au seuil $\alpha=0,001$. On se trouve conduit à ne pas rejeter l'hypothèse d'indépendance pour la relation entre a et b, mais par contre, à rejeter violemment une telle hypothèse d'indépendance pour l'association entre c et d !.

Mais alors imaginons que la situation réelle en ce qui concerne la valeur de $_3\rho(c,d)$ soit celle la plus vraisemblable ; c'est-à-dire $_3\rho(c,d)=0,1$. D'ailleurs, le calcul de l'intervalle de confiance à 0,99 pour $_3\rho(a,b)$ donne $\left[0,088;0,112\right]$ (voir pour ce calcul l'expression (4) ci-dessus).

Dans cette situation la plus vraisemblable, si $_3\rho(a,b)$ était inférieur à $_3\rho(c,d)$ -soit si $_3\rho(a,b)\leqslant 0,1$, la probabilité d'observer pour la statistique $_3R_n(a,b)$, un résultat aussi grand que $_3r_n(a,b)=0,3$, est égale à

$$\Pr\{{}_3R_n(a,b)\geqslant 0,3 \,/\, {}_3\rho(a,b)\leqslant 0,1\}\leqslant \Pr\left\{\frac{{}_3R_n(a,b)-0,1}{\sqrt{\mathrm{var}\left[{}_3R_n(a,b)\right]}}\geqslant\frac{0,2}{\sqrt{0,232/100}}\cong 4,15\right\}\cong 10^{-5}, \qquad (6)$$

la variance de $R_n(a,b)$ étant déterminée à partir de la formule (44) ci-dessus.

Ainsi, dans la situation la plus vraisemblable ($_3\rho(c,d)=0,1$), une hypothèse telle que $_3\rho(a,b)\leqslant{}_3\rho(c,d)$ apparaît comme hautement invraisemblable (cf. (6)) et pourtant -répétons-le- la pratique ci-dessus du test d'indépendance au seuil $\alpha=0,001$, nous fait violemment rejeter l'hypothèse d'indépendance pour la relation entre c et d, mais, ne nous fait pas exclure une telle indépendance entre a et b ! D'où le paradoxe.

Cette dernière circonstance nous renforce dans notre démarche qui consiste à utiliser l'h.a.l. "latérale" (définie de façon plus ou moins combinatoire au niveau de E (cf.§I.1. et V)) pour la construction d'indices normalisés et pour la détermination d'une échelle $\left[0,1\right]$ de fréquence mathématique -destinée à la comparaison deux à deux d'un ensemble de variables descriptives (ici formé d'attributs)- en termes de vraisemblance des liaisons observées (cf.§I.2.).

IV. JUSTIFICATION DE L'ECHELLE DE PROXIMITE POUR LA COMPARAISON DEUX A DEUX D'UN ENSEMBLE DE VARIABLES ; *Distribution de la table d'indices aléatoires.*

La justification de la référence à la loi normale pour obtenir la table (16)(§ I.2.) des indices d'association sur en termes de fréquence mathématique, conduit naturellement à l'étude de la distribution de la table des indices aléatoires d'association, associée à (12) (§I.2.).

Nous commencerons par envisager cette étude dans le cas où les variables sont quantitatives. Cela permettra de mieux situer la nature statistique du problème. D'autre part, de la manière non paramétrique dont il est traité, le cas numérique apparaît (cf. [LERMAN(1981)] Chap.2) comme une généralisation du cas logique, mais seulement dans le cadre de l'h.a.l. N_1.

IV.1. TABLE DES INDICES ALEATOIRES D'ASSOCIATION ENTRE V.A. QUANTITATIVES "NON-PARAMETRIQUES"

Soit $V=\{v_j/1\leqslant j\leqslant m\}$ une suite de m variables numériques quantitatives, respectivement observées sur un ensemble E de n sujets indexé au moyen de $I=\{1,2,\ldots,i,\ldots,n\}$. A la suite de variables observées, l'h.a.l. associe une suite de v.a. $\{v'_j/1\leqslant j\leqslant m\}$ où, pour tout $j=1,\ldots,m$, $v'_j=v_j.\sigma_j$, où $(\sigma_1,\sigma_2,\ldots,\sigma_j,\ldots,\sigma_m)$ est une suite de m permutations aléatoires indépendantes sur $\{1,2,\ldots,i,\ldots,n\}$.

Désignons par μ_j et var(j) la moyenne et la variance empirique de la distribution sur E de la variable v_j :

$$\mu_j = \frac{1}{n} \sum_{1\leqslant i\leqslant n} v_j(i) \text{ et } var(j)= \frac{1}{n} \sum_{1\leqslant i\leqslant n} [v_j(i)-\mu_j]^2 \quad , \quad (1)$$

pour tout $j=1,2,\ldots,m$.

En posant w_j la j-ème variable centrée réduite :

$$w_j=[v_j-\mu_j]/\sqrt{var(j)} \qquad (2)$$

pour tout $j=1,2,\ldots,m$, la statistique aléatoire associée à l'indice normalisé de proximité entre les variables v_j et v_k, peut -en confondant (n-1) et n- se mettre sous la forme

$$Q(w_j.\sigma_j,w_k.\sigma_k)= \frac{1}{\sqrt{n}} \sum_{1\leqslant i\leqslant n} w_j.\sigma_j(i)\ w_k.\sigma_k(i) \qquad (3)$$

La question posée est celle de l'étude de la forme asymptotique de la distribution de la suite de ces indices aléatoires d'association

$$\{Q(w_j.\sigma_j,w_k.\sigma_k)= \frac{1}{\sqrt{n}} \sum_{1\leqslant i\leqslant n} w_j.\sigma_j(i)\ w_k.\sigma_k(i)/1\leqslant j<k\leqslant m\} \qquad (4)$$

Ce problème a un caractère tout à fait fondamental puisqu'il est en quelque sorte le correspondant non paramétrique du vieux problème résolu par J. Wishart [J. WISHART(1928)] dans le cas où les v.a. suivent une loi multinormale. Les résultats que nous obtiendrons ici auront un caractère encore

parcellaire, mais nous l'espérons suffisant pour justifier du bien fondé de la référence à la loi normale utilisée.

Nous allons commencer par déterminer (§ IV.1.1.) la matrice des covariances de la suite des v.a. $\{Q(w'_j,w'_k)/1\leqslant j<k\leqslant m\}$, où nous avons noté w'_j et w'_k pour $w_j.\sigma_j$ et $w_k.\sigma_k$. On verra que cette dernière a une forme très simple.

Au paragraphe IV.1.2., nous admettons une conjecture que nous chercherons à justifier au mieux.

Nous évoquerons ensuite le cas où l'ensemble des variables est un ensemble $\mathcal{Q}$ d'attributs logiques. Pour cette situation, il nous sera possible de préciser une hypothèse d'absence de liaison globale de nature inférentielle pour laquelle la distribution de la table des indices aléatoires $\{Q(a'_j,a'_k)/1\leqslant j<k\leqslant m\}$, est asymptotiquement multinormale.

IV.1.1. Matrice des covariances de la table (4)

Pour déterminer cette matrice des covariances, nous avons besoin de déterminer deux types d'espérances produits : le premier est celui où les deux paires de variables n'ont pas de composante commune et le second, où les deux paires ont une composante commune ; enfin, le dernier cas -qui pour j et k donnés, se réduit à la détermination de la variance de $Q(w'_j,w'_k)$- est celui où les deux paires sont identiques.

IV.1.1.1. Calcul de $\mathscr{E}\left[Q(w'_j,w'_k)Q(w'_h,w'_\ell)\right]$

Précisons une dernière fois que dans l'indexation des lettres différentes indiquent des indices distincts ($\{j,k\}\cap\{h,\ell\}=\emptyset$). Les permutations aléatoires $\sigma_j,\sigma_k,\sigma_h$ et σ_ℓ sont par définition indépendantes. Par conséquent $Q(w_j.\sigma_j,w_k.\sigma_k)$ et $Q(w_h.\sigma_h,w_\ell.\sigma_\ell)$ sont deux v.a. indépendantes. D'autre part, en vertu d'un calcul classique, on a

$$\text{moy.}\left[Q(w_j.\sigma_j,w_k.\sigma_k)\right]=n\ \text{moy.}(w_j)\text{moy.}(w_k)=0 \qquad (5)$$

$$\text{var.}\left[Q(w_j.\sigma_j,w_k.\sigma_k)\right]=\frac{n^2}{(n-1)}\ \text{var.}(w_j)\text{var.}(w_k)=\frac{n}{(n-1)}\cong 1 \qquad (6)$$

Dans ces conditions, l'espérance mathématique exprimée dans le titre est nulle. La covariance entre les deux v.a. $Q(w'_j,w'_k)$ et $Q(w'_h,w'_\ell)$ est nulle.

IV.1.1.2. Calcul de $\mathscr{E}\left[Q(w'_j,w'_k)Q(w'_j,w'_\ell)\right]$

Nous allons voir que dans ce cas également où les deux paires $\{j,k\}$ et $\{j,\ell\}$ ont une composante commune j, les deux v.a. $Q(w'_j,w'_k)$ et $Q(w'_j,w'_\ell)$ sont indépendantes. Fixons en effet la permutation j et sans rien perdre de la généralité, supposons que cette permutation fixée soit celle identique. Les deux v.a. $\sum\{w_j(i)w_k.\sigma_k(i)/1\leqslant i\leqslant n\}$ et $\sum\{w_j(i)w_\ell.\sigma_\ell(i)/1\leqslant i\leqslant n\}$ sont indépendantes en raison de l'indépendance entre σ_k et σ_ℓ. Pour mieux s'en convaincre, nous allons calculer l'espérance mathématique de leur produit :

$$\left(\left[\sum_{1\leq i\leq n} w_j(i)w_k.\sigma_k(i)\right]\left[\sum_{1\leq i'\leq n} w_j(i')w_\ell.\sigma_\ell(i')\right]\right)$$

$$= \frac{1}{(n!)^2}\sum\left\{\left[\sum_i w_j(i)w_k.\sigma_k(i)\right]\left[\sum_{i'} w_j(i')w_\ell.\sigma_\ell(i')\right] \Big/ (\sigma_k,\sigma_\ell)\in G_n \times G_n\right\}, \quad (7)$$

où G_n désigne l'ensemble des n! permutations sur (1,2,...,n).

La dernière somme (21) peut se mettre sous la forme

$$\left(\sum_i w_j(i)\left[\frac{1}{n!}\sum\{w_k.\sigma_k(i)/\sigma_k\in G_n\}\right]\right)\left(\sum_{i'} w_j(i')\left[\frac{1}{n!}\sum\{w_\ell.\sigma_\ell(i')/\sigma_\ell\in G_n\}\right]\right) \quad (8)$$

Or le contenu du premier (resp.second) crochet est égal à moy(w_k) (resp. moy.w) et l'expression (22) peut s'écrire

$$(n\ \text{moy.}(w_j)\ \text{moy.}(w_k))x(n\ \text{moy.}(w_j)\text{moy.}(w_\ell))=0 \quad (9)$$

Il est donc a fortiori vrai que

$$\mathcal{E}\left(\left[\sum_i w_j.\sigma_j(i)w_k.\sigma_k(i)\right]\left[\sum_{i'} w_j.\sigma_j(i')w_\ell.\sigma_\ell(i')\right]\right)$$

$$= (n\ \text{moy.}(w_j)\text{moy.}(w_k))x(n\ \text{moy.}(w_j)\text{moy.}(w_\ell))=0 \quad (10)$$

Ainsi, la covariance entre les deux v.a. $Q(w_j',w_k')$ et $Q(w_j',w_\ell')$ est également nulle.

IV.1.1.3. Calcul de $\mathcal{E}([Q(w_j',w_k')]^2)$

Il s'agit de la variance de $Q(w_j.\sigma_j,w_k.\sigma_k)$ qui est égale à 1.

Dans ces conditions, la matrice des covariances de la table (4) des indices aléatoires est exactement la matrice identité.

D'autre part, comme nous l'avons mentionné au paragraphe II, en vertu d'un célèbre théorème de la Statistique non paramétrique [WALD&WOLFOWITZ(1944), NOETHER(1949),HAJEK(1961)], la loi marginale de $Q(w_j.\sigma_j,w_k.\sigma_k)$ est, pour tout $(j,k):1\leq j<k\leq m$, asymptotiquement normale et centrée réduite dans notre cas (cf. (1) et (2)).

Toutefois, bien entendu, la loi jointe de (4) n'est pas normale, car si les v.a. composantes de (4) sont deux à deux indépendantes, elles ne le sont plus trois à trois comme nous avons pu nous en rendre explicitement compte pour la comparaison deux à deux de trois attributs aléatoires [LERMAN(1984)].

IV.1.2. Conjecture relative à la loi de la somme des carrés des indices aléatoires et justification de la réduction globale des similarités

Nous nous intéressons ici à la statistique définie par la somme des carrés des indices aléatoires d'association :

$$\sum\{Q^2(v'_j,v'_k)/1\leqslant j<k\leqslant m\} \qquad (11)$$

(cf. début du paragraphe IV.1.1. ci-dessus).

Il s'agit dans l'h.a.l. -à caractère permutationel- définie, d'une somme de m(m-1)/2 carrés de v.a. normales centrées réduites, dont d'ailleurs -comme nous venons juste de le voir- la matrice des covariances est la matrice unité.

Conjecture : La distribution asymptotique de probabilité de la somme (11) est une loi du χ^2 à μ degrés de liberté $(\mu=m(m-1)/2)$.

Cette conjecture se trouve vérifiée pour ce qui concerne les deux premiers moments :

PROPRIETE La moyenne et la variance de la statistique (11) sont respectivement égales à μ et à 2μ où $\mu=m(m-1)/2$.

Pour ce qui est de l'espérance mathématique, c'est immédiat puisque sous l'h.a.l. , $Q(v'_j,v'_k)$ est une v.a. (0,1). C'est également clair pour ce qui est de la variance puisque les v.a. $Q(v'_j,v'_k)$ sont deux à deux indépendantes et

$$\mathcal{E}\left[\sum\{Q^2(v'_j,v'_k)/1\leqslant j<k\leqslant m\}\right]^2=\mu\,\mathcal{E}\left[Q^4(v'_j,v'_k)\right]+\mu(\mu-1)\,\mathcal{E}\left[Q^2(v'_j,v'_k)Q^2(v'_{j'},v'_{k'})\right] \quad \text{où}$$

$(j',k')\neq(j,k)$, d'où,

$$= 3\mu+\mu(\mu-1) \qquad (12)$$

Considérons à présent la statistique aléatoire de proximité entre les deux v.a. v'_j et v'_k respectivement associés à v_j et v_k considérés comme faisant partie de l'ensemble V de variables :

$$Q_s(v'_j,v'_k)=\frac{Q(v'_j,v'_k)}{\sqrt{\frac{1}{\mu}\sum\{Q(v'_{j'},v'_{k'})/1\leqslant j'\leqslant k'\leqslant m\}}} \qquad (13)$$

Il y a une très faible liaison entre le numérateur et le dénominateur dont les carrés ont -dans l'h.a.l.- une corrélation de $1/\sqrt{\mu}$. Par conséquent (13) suit une loi de Student, assimilable à une loi normale $\mathcal{N}(0,1)$ (μ grand).

De la sorte, l'expression

$$P(v_j,v_k)=\Phi\left[Q_s(v_j,v_k)\right], \qquad (14)$$

où Φ est toujours la f.r. $\mathcal{N}(0,1)$, se trouve justifiée pour la définition d'une échelle de fréquence mathématique pour la comparaison deux à deux et de façon relative, d'un ensemble de variables descriptives et notamment des attributs que nous allons aussitôt reprendre.

IV.2. DISTRIBUTION DE LA TABLE DES INDICES D'ASSOCIATION ENTRE ATTRIBUTS ALEATOIRES SOUS LES H.A.L. N_1,N_2 ET N_3

La place nous manque ici de détailler cette étude. Signalons que dans le cas où m=3, nous avons très exactement déterminé cette distribution pour chacune des h.a.l. N_1,N_2 et N_3. D'autre part, dans le cas de m quelconque, nous avons dans chacun des cas établi la matrice des covariances de la suite ordonnée des indices aléatoires d'association dont la forme est particulière et que nous démontrons être définie [LERMAN(1984)].

Rappelons que N_1 est de même nature que l'h.a.l. à caractère permutationel considérée ci-dessus. De sorte que tout ce qui a été exprimé au paragraphe IV.1. ci-dessus s'applique tel quel dans le cas de la table des indices d'association entre attributs aléatoires sous l'h.a.l. N_1 et donc la justification -qui repose sur une conjecture- de la table (16)(§ I.2.) des indices d'association en termes de vraisemblance des liens.

IV.3. H.A.L. COMPATIBLE AVEC N_2 PUIS AVEC N_3 OU LA DISTRIBUTION DE LA MATRICE DES INDICES ALEATOIRES D'ASSOCIATION EST ASYMPTOTIQUEMENT MULTINORMALE

IV.3.1. Expression de l'h.a.l. N_{02} et de la suite des v.a. $\{n(a'_j \wedge a'_k)/1 \leqslant j < k \leqslant m\}$.

En codant par 1(resp. 0), la présence (resp.absence) d'un même attribut, l'h.a.l. N_{02} munit le cube $\{0,1\}^m$ d'une mesure de probabilité produit : si $(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_j,\ldots,\varepsilon_m)$ est un point de $\{0,1\}^m$,

$$p(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_j,\ldots,\varepsilon_m) = \prod_{1\leqslant j\leqslant m} p(\varepsilon_j), \qquad (15)$$

où $p(\varepsilon_j)=\alpha_j$(resp.$\bar{\alpha}_j$) où α_j(resp.$\bar{\alpha}_j$) est la proportion définie au niveau de la population $\mathcal{P}$ des individus possédant l'attribut a_j(resp.$\bar{a}_j$).

$(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_j,\ldots,\varepsilon_m)$ code un attribut croisé multiple et on se trouve devant un modèle d'urne 2^m-nomiale où à l'échantillon observé E, on associe un échantillon aléatoire $\mathcal{E}$ formé d'une suite de n individus aléatoires indépendants où la probabilité pour un même individu de posséder l'attribut croisé $\varepsilon_1 \wedge \varepsilon_2 \wedge \ldots \wedge \varepsilon_j \wedge \ldots \wedge \varepsilon_m$, est définie par le second membre de (53).$n(\varepsilon'_1 \wedge \varepsilon'_2 \wedge \ldots \wedge \varepsilon'_m)$ désignera le nombre aléatoire associé au nombre observé $n(\varepsilon_1 \wedge \varepsilon_2 \wedge \ldots \wedge \varepsilon_m)$ de sujets possédant $\varepsilon_1 \wedge \varepsilon_2 \wedge \ldots \wedge \varepsilon_m$. La loi jointe de $\{n(\varepsilon'_1 \wedge \varepsilon'_2 \wedge \ldots \wedge \varepsilon'_m)/(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_m) \in \{0,1\}^m\}$ est une loi 2^m-nomiale dont les paramètres sont n et les probabilités (15).

Pour fixer les idées, mais sans pour cela restreindre en rien la généralité, nous allons effectuer les écritures pour m=4. Nous commencerons par exprimer la suite des v.a. qui nous intéresse

$$\{n(a'_j \wedge a'_k)/1 \leqslant j < k \leqslant 4\} \qquad (16)$$

par rapport à celle que nous venons d'introduire

$$\{n(\varepsilon'_1 \wedge \varepsilon'_2 \wedge \varepsilon'_3 \wedge \varepsilon'_4)/(\varepsilon_1,\varepsilon_2,\varepsilon_3,\varepsilon_4) \in \{0,1\}^4\} \qquad (17)$$

Pour simplifier ici les notations, désignons par X_{jk} la v.a. $n(a'_j, a'_k)$, $1 \leqslant j < k \leqslant 4$, et par $X_{\varepsilon_1\varepsilon_2\varepsilon_3\varepsilon_4}$ celle $n(\varepsilon'_1 \wedge \varepsilon'_2 \wedge \varepsilon'_3 \wedge \varepsilon'_4), (\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4) \in \{0,1\}^4$. On a les relations

$$\left.\begin{aligned} X_{12} &= X_{1100} + X_{1101} + X_{1110} + X_{1111} \\ X_{13} &= X_{1010} + X_{1011} + X_{1110} + X_{1111} \\ X_{14} &= X_{1001} + X_{1011} + X_{1101} + X_{1111} \\ X_{23} &= X_{0110} + X_{0111} + X_{1110} + X_{1111} \\ X_{24} &= X_{0101} + X_{0111} + X_{1101} + X_{1111} \\ X_{34} &= X_{0011} + X_{0111} + X_{1011} + X_{1111} \end{aligned}\right\} \qquad (18)$$

qui restent valables lorsqu'on centre les différentes v.a. considérées de la forme X_{jk} ou $X_{\varepsilon_1\varepsilon_2\varepsilon_3\varepsilon_4}$.

On peut remarquer que pour l'expression d'un même X_{jk}, les deux composantes indiciaires ε_j et ε_k sont obligatoirement égales à 1, de sorte que le vecteur colonne $t(X_{12}, X_{13}, X_{14}, X_{23}, X_{24}, X_{34})$ s'exprime linéairement par rapport à la suite partielle des $X_{\varepsilon_1\varepsilon_2\varepsilon_3\varepsilon_4}$ -où deux composantes indiciaires au moins sont égales à 1- et que nous rangeons lexicographiquement : $X_{0011}, X_{0101}, X_{0110}, X_{0111}, X_{1001}, X_{1010}, X_{1011}, X_{1100}, X_{1101}, X_{1110}, X_{1111}$.

Ce dernier vecteur colonne à 11 composantes, correspondant à un vecteur multinomial tronqué, suit asymptotiquement une loi multinormale dont la matrice des covariances est définie.

La matrice de la transformation linéaire permettant de passer de ce dernier vecteur à celui $t(X_{12}, X_{13}, X_{14}, X_{23}, X_{24}, X_{34})$ qui nous intéresse est

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad (19)$$

et qui est très manifestement de rang $6 = \binom{4}{2}$.

Plus généralement, cette matrice est de dimension $\binom{m}{2} \times (2^m - 1 - m)$ et de rang $\binom{m}{2}$.

Dans ces conditions, le vecteur ${}^t(X_{12}, X_{13}, X_{14}, X_{23}, X_{24}, X_{34})$ suit asymptotiquement une loi multinormale dont la matrice des covariances -que nous allons directement calculer- est définie.

IV.3.2. Calcul dans le cadre de N_{02} de la matrice des covariances de $\{n(a'_j,a'_k)/1\leqslant j<k\leqslant m\}$.

Nous allons continuer à effectuer nos écritures dans le cas -non restrictif pour la généralité- où m=4 et nous allons désigner par {a,b,c,d} l'ensemble des quatre attributs. Trois calculs structurellement distincts sont à considérer : $\mathrm{var}[n(a'\wedge b')]$, $\mathrm{cov}.[n(a'\wedge b'), n(a'\wedge c')]$ et $\mathrm{cov}[n(a'\wedge b'),n(c'\wedge d')]$. Ces calculs se réfèrent -dans le cadre de N_{02}- aux décompositions suivantes :

$$\left.\begin{array}{l} n(a'\wedge b')=n(a'\wedge b'\wedge c'\wedge d')+n(a'\wedge b'\wedge c'\wedge \bar{d}')+n(a'\wedge b'\wedge \bar{c}'\wedge d')+n(a'\wedge b'\wedge \bar{c}'\wedge \bar{d}') \\ n(a'\wedge c')=n(a'\wedge b'\wedge c'\wedge d')+n(a'\wedge \bar{b}'\wedge c'\wedge \bar{d}')+n(\bar{a}'\wedge \bar{b}'\wedge c'\wedge d')+n(\bar{a}'\wedge \bar{b}'\wedge c'\wedge \bar{d}') \\ n(c'\wedge d')=n(a'\wedge b'\wedge c'\wedge d')+n(a'\wedge \bar{b}'\wedge c'\wedge d')+n(\bar{a}'\wedge b'\wedge c'\wedge d')+n(\bar{a}'\wedge \bar{b}'\wedge c'\wedge d') \end{array}\right\} \quad (20)$$

On a, après développement et en désignant par α,β,γ et δ, les proportions définies au niveau de la population parente $\mathcal{P}$, des individus possédant, respectivement, les attributs a,b,c et d,

$$\mathrm{var}.[n(a'\wedge b')]=n\alpha\beta\gamma\delta(1-\alpha\beta\gamma\delta)+n\alpha\beta\gamma\bar{\delta}(1-\alpha\beta\gamma\bar{\delta})+n\alpha\beta\bar{\gamma}\delta(1-\alpha\beta\bar{\gamma}\delta)+n\alpha\beta\bar{\gamma}\bar{\delta}(1-\alpha\beta\bar{\gamma}\bar{\delta})$$
$$-2n[\alpha\beta\gamma\delta\alpha\beta\gamma\bar{\delta}+\alpha\beta\gamma\delta\alpha\beta\bar{\gamma}\delta+\alpha\beta\gamma\delta\alpha\beta\bar{\gamma}\bar{\delta}+\alpha\beta\gamma\bar{\delta}\alpha\beta\bar{\gamma}\delta+\alpha\beta\gamma\bar{\delta}\alpha\beta\bar{\gamma}\bar{\delta}+\alpha\beta\bar{\gamma}\delta\alpha\beta\bar{\gamma}\bar{\delta}]$$

où $\bar{\alpha}=(1-\alpha)$, $\bar{\beta}=(1-\beta)$, $\bar{\gamma}=(1-\gamma)$ et $\bar{\delta}=(1-\delta)$.

$$\mathrm{var}.[n(a',b')]=n\alpha\beta x\{1-\alpha\beta[1+2(\gamma\bar{\gamma}(1-\delta\bar{\delta})+\delta\bar{\delta}(1-\gamma\bar{\gamma}))]\} \quad (21)$$

$$\mathrm{cov}.[n(a',b'),n(a',c')]=n\alpha\beta\gamma\delta\{(1-\alpha\beta\gamma\delta)-\alpha\beta\gamma\bar{\delta}-\alpha\beta\bar{\gamma}\delta-\alpha\beta\bar{\gamma}\bar{\delta}\}$$
$$+n\alpha\beta\gamma\bar{\delta}\{-\alpha\beta\gamma\delta+(1-\alpha\beta\gamma\bar{\delta})-\alpha\beta\gamma\delta-\alpha\beta\bar{\gamma}\delta-\alpha\beta\bar{\gamma}\bar{\delta}\}$$
$$+n\alpha\beta\bar{\gamma}\delta\{-\alpha\beta\gamma\delta-\alpha\beta\gamma\bar{\delta}-\alpha\beta\bar{\gamma}\delta-\alpha\beta\bar{\gamma}\bar{\delta}\}$$
$$+n\alpha\beta\bar{\gamma}\bar{\delta}\{-\alpha\beta\gamma\delta-\alpha\beta\gamma\bar{\delta}-\alpha\beta\bar{\gamma}\delta-\alpha\beta\bar{\gamma}\bar{\delta}\}$$

$$=n\alpha\beta\gamma-n\alpha\beta\{\alpha\gamma\}$$
$$=n\alpha\bar{\alpha}\beta\gamma \quad (22)$$

Il reste maintenant à déterminer $\mathrm{cov}[n(a',b'),n(c',d')]$. On a

$$\mathrm{cov}.[n(a'\wedge b'),n(c'\wedge d')]=n\alpha\beta\gamma\delta\{1-\alpha\beta\gamma\delta)-\alpha\bar{\beta}\gamma\delta-\bar{\alpha}\beta\gamma\delta-\bar{\alpha}\bar{\beta}\gamma\delta\}$$
$$+n\alpha\beta\gamma\bar{\delta}\{-\alpha\beta\gamma\delta-\alpha\bar{\beta}\gamma\delta-\bar{\alpha}\beta\gamma\delta-\bar{\alpha}\bar{\beta}\gamma\delta\}$$
$$+n\alpha\beta\bar{\gamma}\delta\{-\alpha\beta\gamma\delta-\alpha\bar{\beta}\gamma\delta-\bar{\alpha}\beta\gamma\delta-\bar{\alpha}\bar{\beta}\gamma\delta\}$$
$$+n\alpha\beta\bar{\gamma}\bar{\delta}\{-\alpha\beta\gamma\delta-\alpha\bar{\beta}\gamma\delta-\bar{\alpha}\beta\gamma\delta-\bar{\alpha}\bar{\beta}\gamma\delta\}$$

$$=n\alpha\beta\gamma\delta-n\alpha\beta\gamma\delta$$
$$=0 \quad (23)$$

THEOREME. Sous l'hypothèse d'absence de liaison N_{02} définie ci-dessus, la distribution asymptotique de la suite des v.a. $\{n(a'_j\wedge a'_k)/1\leqslant j<k\leqslant m\}$ est asymptotiquement normale de matrice des covariances -déterminée par les formules (21),(22) et (23)- définie.

IV.3.3. Matrice des covariances dans le cas de l'h.a.l. N_{03} correspondant à un modèle Poissonnien

Dans le cadre d'une telle hypothèse, la loi jointe de $\{n(\varepsilon_1' \varepsilon_2' \dots \varepsilon_m')/(\varepsilon_1,\dots,\varepsilon_2,\varepsilon_m) \in \{0,1\}^m\}$ est un produit de 2^m lois indépendantes de Poisson, où le paramètre de la loi de $n(\varepsilon_1' \varepsilon_2' \dots \varepsilon_m')$ est $n\Pi\{\alpha_j^{\varepsilon_j}(\bar{\alpha}_j)^{(1-\varepsilon_j)} / 1\leq j\leq m\}$.

Comme ci-dessus, pour fixer les idées mais sans restreindre la généralité, revenons au cas où m=4. Reportons-nous dans ces conditions aux décompositions (58) qui nous permettent d'obtenir

$$\begin{array}{ll} \text{var.} & \{n(a'\wedge b')\} = n\alpha\beta \\ \text{cov.} & \{n(a'\wedge b'), n(a'\wedge c')\} = n\alpha\beta\gamma\delta + n\alpha\beta\gamma\bar{\delta} = n\alpha\beta\gamma \\ \text{cov.} & \{n(a'\wedge b'), n(c'\wedge d')\} = n\alpha\beta\gamma\delta \end{array} \qquad (24)$$

Ainsi, la matrice des variances et covariances de la suite $(n(a'\wedge b'), n(a'\wedge c'), n(a'\wedge d'), n(b'\wedge c'), n(b'\wedge d'), n(c'\wedge d'))$, se met sous la forme

$$\alpha\beta \begin{bmatrix} 1 & \gamma & \delta & \gamma & \delta & \gamma\delta \\ \beta & 1 & \delta & \beta & \beta\delta & \delta \\ \beta & \gamma & 1 & \beta\gamma & \beta & \gamma \\ \alpha & \alpha & \alpha\beta & 1 & \delta & \delta \\ \alpha & \alpha\gamma & \alpha & \gamma & 1 & \gamma \\ \alpha\beta & \alpha & \alpha & \beta & \beta & 1 \end{bmatrix} \qquad (25)$$

On peut aisément vérifier que cette matrice est de rang 6. En effet, le système des vecteurs lignes est trivialement équivalent à celui obtenu par substitution linéaire :

$$\left[\begin{array}{c|c|c|c|c|c} 1 & \gamma & \delta & \gamma & \delta & \gamma\delta \\ 0 & 1-\beta\gamma & \delta(1-\beta) & \beta(1-\gamma) & 0 & \delta(1-\beta\gamma) \\ 0 & \gamma(1-\beta) & 1-\beta\delta & 0 & \beta(1-\delta) & \gamma(1-\beta\delta) \\ 0 & \alpha(1-\gamma) & 0 & 1-\alpha\gamma & \delta(1-\alpha) & \delta(1-\alpha\gamma) \\ 0 & 0 & \alpha(1-\delta) & \gamma(1-\alpha) & 1-\alpha\delta & \gamma(1-\alpha\delta) \\ 0 & \alpha(1-\alpha\gamma) & \alpha(1-\delta) & \beta(1-\alpha\gamma) & \beta(1-\alpha\delta) & 1-\alpha\beta\gamma\delta \end{array}\right] \qquad (26)$$

où le premier vecteur est indépendant de la suite des autres qui forme un système libre. De façon générale, on a le résultat suivant :

THEOREME. Sous l'hypothèse d'absence de liaison N_{03} du modèle Poissonnien, la distribution asymptotique de la suite des v.a. $\{n(a_j'\wedge a_k')/1\leq j<k\leq m\}$ est asymptotiquement normale de matrice des covariances -déterminée par les formules (24)- définie

IV.3.4. Sur deux modes nouveaux de réduction globale des similarités

Ces deux modes sont une conséquence directe de l'analyse des paragraphes précédents (IV.3.1. , IV.3.2. et IV.3.3.). Désignons par V_{02} (resp. V_{03}) la

matrice des covariances obtenue au paragraphe IV.2. (resp. IV.3.) et par $q'={}^t(q(a'_j,a'_k)/1\leqslant j<k\leqslant m)$, le vecteur colonne des indices aléatoires "centrés" de proximité $(q(a'_j,a'_k)=[n(a'_j,a'_k)-n\alpha_j\alpha_k])$, on déduit des théorèmes précédents que sous l'hypothèse N_{02}(resp. N_{03}) la v.a. $tq'V_{02}q'$(resp. ${}^tq'V_{03}q'$) suit une loi du χ^2 à $\mu=\binom{m}{2}$ degrés de liberté [LANCASTER(1969)].

Si maintenant on désigne par $V_{01\varepsilon}$ (ε=1,2 ou 3) la matrice diagonale des variances $\{var_\varepsilon[q(a'_j,a'_k)]/1\leqslant j<k\leqslant m\}$ où var_ε est la variance calculée dans l'h.a.l. N_ε, la formule $\left[{}^tq\,V_{01\varepsilon}q/\binom{m}{2}\right]^{1/2}$ (13) (§ IV.1.) propose une réduction globale au moyen de $\left[{}^tq\,V_{01\varepsilon}q/\binom{m}{2}\right]^{1/2}$ qu'on justifie au mieux au moyen de la conjecture (pour ε=1) du paragraphe IV.1. Suite à la précédente analyse, nous pouvons proposer deux autres modes parfaitement justifiés de réduction globale ; le premier au moyen de $\left[{}^tqV_{02}q/\binom{m}{2}\right]^{1/2}$ et le second, au moyen de $\left[{}^tqV_{03}q/\binom{m}{2}\right]^{1/2}$.

V. CONCLUSION : *situations respectives de notre approche et de celle de GOODMAN et KRUSKAL*

Revenons ici sur notre démarche générale dans l'élaboration d'un coefficient d'association entre deux variables de description statistique. Cette dernière qui tire son origine dans les travaux de K. Pearson, M.G. Kendall, A. Wald et J. Wolfowitz, peut être schématisée par le diagramme suivant :

$$(\alpha,\beta)\in AxB \longrightarrow (R(\alpha),R(\beta)) \in \Omega \times \Omega \qquad (1)$$

$$s=card[R(\alpha)\cap R(\beta)] \qquad (2)$$

h.a.l. "respectant les caractéristiques de cardinalité de α et de β : N."

$$S=card[R(\alpha')\cap R(\beta')] \qquad (3)$$

$$Q(\alpha,\beta)=[s-\mathscr{E}(S)]/\sigma(S) \qquad (4)$$

$$P(\alpha,\beta)=Pr\{S\leqslant s/N\}\cong\Phi[Q(\alpha,\beta)] \qquad (5)$$

Figure 1.

que nous allons illustrer dans deux situations classiques.

La première est celle de la comparaison de deux variables qualitatives nominales ; de sorte que α et β sont deux partitions où nous désignons par $t(\alpha)$[resp.$t(\beta)$] le type de la partition α (resp.β) ; c'est-à-dire la suite ordonnée des cardinaux de ses classes. Dans ces conditions A(resp.B) est l'ensemble des partitions sur E de type $t(\alpha)$[resp.$t(\beta)$]. $R(\alpha)$[resp.$R(\beta)$] est l'ensemble des paires sous-ensemble de l'ensemble $P_2(E)$ des parties à deux éléments de E dont les deux composantes sont réunies dans une même classe de la partition α(resp.β). Ω peut être défini comme l'ensemble des parties de $P_2(E)$ dont chacune correspond à la représentation d'une relation d'équivalence sur E.

De façon plus explicite, notons

$\alpha=\{E_i/1\leqslant i\leqslant I\}$, $\beta=\{F_j/1\leqslant j\leqslant J\}$,

$t(\alpha)=\{n_{i.}/1\leqslant i\leqslant I\}$, $t(\beta)=\{n_{.j}/1\leqslant j\leqslant J\}$,

$p(\alpha)=\{p_{i.}=n_{i.}/n \,/\, i\leqslant i\leqslant I\}$, $p(\beta)=\{p_{.j}=n_{.j}/n \,/\, 1\leqslant j\leqslant J\}$,

$$\alpha\wedge\beta=\{E_i\cap F_j/1\leqslant i\leqslant I,\ 1\leqslant j\leqslant J\}, \qquad (1)$$

$t(\alpha\wedge\beta)=\{n_{ij}=\mathrm{card}(E_i\cap F_j/1\leqslant i\leqslant I,\ 1\leqslant j\leqslant J\}$,

et

$p(\alpha\wedge\beta)=\{p_{ij}=n_{ij}/n \,/\, 1\leqslant i\leqslant I,\ 1\leqslant i\leqslant J\}$ où

$\alpha\wedge\beta$ est la partition croisée dont $t(\alpha\wedge\beta)$ définit la table de contingence et $p(\alpha\wedge\beta)$, la table des proportions ($n=\mathrm{card}(E)$).

En représentant une partition par l'ensemble des paires qu'elle réunit, l'indice brut se met sous la forme :

$$s=\mathrm{card}\left[R(\alpha)\cap R(\beta)\right]=\mathrm{card}\left[R(\alpha\wedge\beta)\right]=\sum\{n_{ij}(n_{ij}-1)/2 \,/\, 1\leqslant i\leqslant I,\ 1\leqslant j\leqslant J\} \qquad (2)$$

L'espérance mathématique $\mathcal{E}[s(\alpha',\beta')]$ et la variance $\mathrm{var}[s(\alpha',\beta')]$ ont respectivement pour formes [LERMAN(1973),(1981)] :

$\lambda\mu$ et $\lambda\mu + \rho\sigma + (\theta\zeta - \lambda^2\mu^2)$, où

$$\left.\begin{aligned}\lambda&=\sum_{1\leqslant i\leqslant I} n_{i.}(n_{i.}-1)/\sqrt{2n(n-1)},\quad \rho=\sum_{1\leqslant i\leqslant I} n_{i.}(n_{i.}-1)(n_{i.}-2)/\sqrt{n(n-1)(n-2)},\\ \theta&=\{\Big[\sum_{1\leqslant i\leqslant I} n_{i.}(n_{i.}-1)\Big]^2-2\sum_{1\leqslant i\leqslant I} n_{i.}(n_{i.}-1)(2n_{i.}-3)\}/2\sqrt{n(n-1)(n-2)(n-3)}\end{aligned}\right\}\text{ et où} \qquad (3)$$

les expressions de μ, σ et ζ ont respectivement la même forme que λ, ρ et θ ; les $n_{i.}$ de $t(\alpha)$ étant remplacés par les $n_{.j}$ de $t(\beta)$, $1\leqslant i\leqslant I$, $1\leqslant j\leqslant J$. D'où l'expression de l'indice $Q(\alpha,\beta)$.

Imaginons à présent (cas usuel) que les n_{ij} soient "assez grands", on obtient LERMAN(1984) , que la forme limite de l'indice d'association $Q(\alpha,\beta)$ est -au facteur 1/2 près-

$$\frac{\sqrt{n}\ \times\ \sum\{(p_{ij}^2-p_{i.}^2\,p_{.j}^2)/1\leqslant i\leqslant I,\ 1\leqslant j\leqslant J\}}{\sqrt{\Big[(\sum_i p_{i.}^2)^2-(\sum_i p_{i.}^3)\Big]\Big[(\sum_j p_{.j}^2)^2-(\sum_j p_{.j}^3)\Big]}} \qquad (4)$$

Il est remarquable de constater que comme dans le cas de la comparaison des attributs de description (cf. coefficient de K. Pearson), c'est le même facteur $\sqrt{n}$ qui apparaît avant une fonction $f[p(\alpha\wedge\beta)]$ du tableau des proportions $p(\alpha\wedge\beta)$ (cf. (1) ci-dessus). Cette fonction f définit parfaitement un indice qui peut être appliqué au tableau des mêmes proportions $\pi(\alpha\wedge\beta)$, mais considéré au niveau de la population entière $\mathcal{P}$.

L'étude statistique mentionnée ci-dessus a donc été essentielle pour la découverte de l'expression formelle de l'indice (4).

Un autre indice classique et bien connu pour cette même situation, est celui de A.A. Tschuprow [TSCHUPROW(1934)] qui se met sous la forme suivante :

$$T_{\alpha\beta} = \phi^2_{\alpha\beta}/\sqrt{(I-1)(J-1)}, \qquad (5)$$

où

$$\phi^2_{\alpha\beta} = \sum\{(p^2_{ij}/p_{i.}p_{.j})\Big/1\leq i\leq I,\ 1\leq j\leq J\}-1, \quad (6)$$

$$= \chi^2(\alpha,\beta)/n. \qquad (7)$$

L'indice ϕ^2 ou $T_{\alpha\beta}$ est de nature différente de celui se déduisant de (4) ci-dessus (en divisant par $\sqrt{n}$). Néanmoins, on peut considérer que son expression formelle se justifie par une étude statistique préalable au niveau de E, puisqu'on peut montrer que $\chi^2(\alpha',\beta')$ -où (α',β') est un couple de partitions aléatoires indépendantes de types respectifs $t(\alpha)$ et $t(\beta)$- suit une loi du χ^2 à $(I-1)(J-1)$ degrés de liberté.

Considérons à présent la deuxième situation classique où les deux variables qualitatives α et β sont ordinales (i.e. l'ensemble des modalités d'une même variable est totalement ordonné). α(resp. β) définit un préordre total sur l'ensemble E des individus dont les classes sont les E_i(resp.F_j), $1\leq i\leq I$ (resp.$1\leq j\leq J$) (cf.(1)). On considère

$$R(\alpha)= \sum\{E_i xE_{i'}/1\leq i<i'\leq I\},\ R(\beta)= \sum\{F_j xF_{j'}/1\leq j<j'\leq J\ , \quad (8)$$

de sorte que l'indice brut se met sous la forme

$$s=\mathrm{card}\left[R(\alpha)\cap R(\beta)\right]= \sum\{n_{ij}n_{i'j'}/1\leq i<i'\leq I,\ 1\leq j<j'\leq J\}. \quad (9)$$

L'espérance mathématique $\mathcal{E}[s(\alpha',\beta')]$ et la variance var.$[s(\alpha',\beta')]$ s'écrivent respectivement [LERMAN(1973),(1981),(1983c)] :

$$\lambda\mu \text{ et } [\lambda\mu+\rho_{cc}\sigma_{cc}+\rho_{ff}\sigma_{ff}+2\rho_{cf}\sigma_{cf}+(\theta\zeta-\lambda^2\mu^2)]. \quad (10)$$

Les expressions de $\mu,\sigma_{cc},\sigma_{ff},\sigma_{cf}$ et ζ sont respectivement de même forme que celles $\lambda,\rho_{cc},\rho_{ff},\rho_{cf}$ et θ ; si les premières sont relatives à la composition $t(\alpha)=\{n_{i.}/1\leq i\leq I\}$ du préordre total associé à la variable α, les secondes sont relatives à la composition $t(\beta)=\{n_{.j}/1\leq j\leq J\}$ du préordre total associé à la variable β. Plus précisément

$$\lambda = \frac{1}{\sqrt{n(n-1)}} \sum \{n_{i.} n_{i'.} / 1 \leqslant i < i' \leqslant I\}$$

$$\rho_{cc} = \frac{1}{\sqrt{n(n-1)(n-2)}} \sum \{n_{i.} n^{c}_{i.} (n^{c}_{i.} - 1)/2 \leqslant i \leqslant I\}$$

$$\rho_{ff} = \frac{1}{\sqrt{n(n-1)(n-2)}} \sum \{n_{i.} n^{f}_{i.} (n^{f}_{i.} - 1)/2 \leqslant i \leqslant I\} \qquad (11)$$

$$\rho_{cf} = \frac{1}{\sqrt{n(n-1)(n-2)}} \sum \{n_{i.} n^{c}_{i.} n^{f}_{i.} / 2 \leqslant i \leqslant (I-1)\}$$

$$\theta = \frac{1}{\sqrt{n(n-1)(n-2)(n-3)}} \sum \{n_{i.} n_{i'.} [\{\sum \{n_{\ell.} n_{\ell'.} / 1 \leqslant \ell < \ell' \leqslant I\} + n_{i.} + n_{i'.} - 2n + 1] /$$

où on note $1 \leqslant i < i' \leqslant I\}$,

$$n^{c}_{i.} = \sum \{n_{\ell.} / \ell < i \quad \text{et} \quad n^{f}_{i.} = \sum \{n_{\ell.} / \ell > i\}$$

Le calcul de la forme asymptotique de l'indice $Q(\alpha,\beta)$ [LERMAN(1984)], montre que le numérateur du rapport qui le définit se présente comme suit :

$$\sqrt{n} \sum \{(p_{ij} p_{i'j'} - p_{i.} p_{i'.} p_{.j} p_{.j'}) / 1 \leqslant i < i' \leqslant I,\ 1 \leqslant j < j' \leqslant J\}, \qquad (12)$$

et le dénominateur représente la racine carrée de

$$4\Big(\sum_{i<i'} p_{i.} p_{i'.}\Big)^2 \Big(\sum_{j<j'} p_{.j} p_{.j'}\Big)^2 + \Big[\sum_{i} p_{i.} (p^{c}_{i.})^2\Big] \Big[\sum_{j} p_{.j} (p^{c}_{.j})^2\Big]$$

$$+ \Big[\sum_{i} p_{i.} (p^{f}_{i.})^2\Big] \Big[\sum_{j} p_{.j} (p^{f}_{.j})^2\Big]$$

$$+ 2\Big(\sum_{i} p_{i.} p^{c}_{i.} p^{f}_{i.}\Big) \Big(\sum_{j} p_{.j} p^{c}_{.j} p^{f}_{.j}\Big) - \Big(\sum_{i<i'} p_{i.} p_{i'.}\Big)^2 \Big[\sum_{j<j'} p_{.j} (1 - p_{.j}) p_{.j'}$$

$$+ \sum_{j<j'} p_{.j} p_{.j'} (1 - p_{.j'})\Big]$$

$$- \Big(\sum_{j<j'} p_{.j} p_{.j'}\Big)^2 \Big[\sum_{i<i'} p_{i.} (1 - p_{i.}) p_{i'.} + \sum_{i<i'} p_{i.} p_{i'.} (1 - p_{i'.})\Big] \qquad (13)$$

On remarquera que -comme pour la formule (4) ci-dessus de comparaison de deux variables "partition"- l'expression limite de l'indice d'association entre deux variables "préordre total", se présente comme $\sqrt{n}\, \mathcal{G}[p(\alpha \wedge \beta)]$, où $\mathcal{G}[p(\alpha \wedge \beta)]$ est la fonction du tableau $p(\alpha \wedge \beta)$ des proportions $p_{ij}, 1 \leqslant i \leqslant I, 1 \leqslant j \leqslant J$ (cf.(1) . Cette fonction $\mathcal{G}$ définit parfaitement un indice qui peut être appliqué au tableau des mêmes proportions $\pi(\alpha \wedge \beta)$, mais considéré au niveau de la population totale .

Or les indices que proposent Goodman et Kruskal [GOODMAN & KRUSKAL(1954)] -et notamment le coefficient γ- n'obéissant pas à notre forme d'analyse statistique préalable au niveau intrinsèque de E (sans référence à une population parente $\mathcal{P}$, ce qui aurait assuré leur cohérence formelle et statistique. En effet, l'expression formelle de chacun de ces indices est, à partir d'une intuition première, posée a priori, donc non sans arbitraire. Nous avons d'ailleurs pu voir [LERMAN(1973),(1981)] qu'à notre sens, il y avait un biais statistique dans l'expression de l'indice d'association entre deux variables qualitatives ordinales, telle qu'elle se trouvait posée par M.G. Kendall qui la déduisait de l'algorithme de calcul du coefficient τ de comparaison de deux variables "rang" [KENDALL(1970)]. Ce type de biais existe également pour l'indice γ.

REFERENCES

M. ALLAIS ; *"Fréquence, probabilité et hasard"*, Journal de la Société de Statistique de Paris, n°2 - 2ème trimestre (1983).

M. BLANCARD ; *"Analyse d'un important fichier de bilans de santé"*, rapport de DEA, Univ. de Rennes I, Sept.(1976).

F. BONNIEUX, P. RAINELLI, T. CHANTREL et I.C. LERMAN ; *"Construction d'indicateurs socio-économiques liés à la qualité de l'eau"*, Colloque international I.R.I.A. "Analyse des Données et Informatique", Versailles, Sept.(1977).

L.A. GOODMAN and W.H. KRUSKAL ; *"Measures of association for cross classifications"*, J.A.S.A. 49, Dec.(1954), 732-764.

L.A. GOODMAN and W.H. KRUSKAL ; *"Measures of association for cross classifications, approximate sampling theory"*, J.A.S.A. 58, June(1963), 310-364.

L.A. GOODMAN and W.H. KRUSKAL ; *"Measures of association for cross classifications, IV : simplification of asumptotic variances"*, J.A.S.A. 67, June(1972), 415-421.

J. HAJEK ; *"Some extensions of the Wald-Wolfowitz-Noether theorem"*, Ann. Math. Stat. 32, (1961), 506-523.

M.G. KENDALL ; *"Rank correlation methods"*, London, Charles Griffin, fourth edition,(1970).

H.O. LANCACTER ; *"The chi-squared distribution"*, John Wiley,(1969).

I.C. LERMAN ; *"Les bases de la classification automatique"*, Paris, Gauthier-Villars, collection Programmation,(1970).

I.C. LERMAN ; *"Introduction à une méthode de classification automatique illustrée par la recherche d'une typologie de personnages enfants à travers la littérature enfantine"*, Revue de Statistique Appliquée, vol.XXI n°3 pp. 23-49, (1973a).

I.C. LERMAN ; *"Etude distributionnelle de statistiques de proximité entre structures finies de même type ; Application à la classification automatique"*, Paris, cahiers du B.U.R.O. 19, 1-52 (1973).

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

I.C. LERMAN ; *"Indices d'association partielle entre variables qualitatives "nominales""*, R.A.I.R.O. 17 n°3, 213-259, Août(1983).

I.C. LERMAN ; *"Indices d'association partielle entre variables "qualitatives" ordinales"*, Pub. Inst. Stat. Univ., XXVIII, fasc. 1,2, p.7-46, (1983).

I.C. LERMAN ; *"Association entre variables qualitatives ordinales "nettes" ou "floues"*, Pub. Int. Rennes IRISA, n°191, Mars(1983).

I.C. LERMAN ; *"Interprétation non linéaire d'un coefficient d'association entre modalités d'une juxtaposition de tables de contingence"*, Rev. Math. Sc. Hum. (21è année, n°83), 5-30,(1983).

I.C. LERMAN ; *"Justification et validité statistique d'une échelle 0,1 de structure de proximité sur un ensemble de variables observées"*. Rap. Int. IRISA, 47 pages, Janv.(1984).

M.H. NICOLAU ; *"Analyse d'un algorithme de classification"*, Thèse de 3ème cycle, Univ. Paris VI, ISUP,(1972).

G.E. NOETHER ; *"On a theorem by Wald and Wolfowitz"*, Ann. Math. Stat. 20, 455-458,(1949).

A.A. TSCHUPROW ; *"The mathematical foundations of the methods to be used in statistical investigation of the dependance between two chance variables"*, Nordisk Statistik Tidskrift, 5,34,(1934).

A. WALD and J. WOLFOWITZ ; *"Statistical tests based on permutations of the observations"*, Ann. Math. Stat. 15, 358-372,(1944).

J. WISHART ; *"The generalized product moment distribution in samples from a normal multivariate population"*, Biometrika, Vol. 20A, 32-52,(1928).

Reçu en Octobre 1984