MATHÉMATIQUES ET SCIENCES HUMAINES

ISRAËL-CÉSAR LERMAN

Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles. 1ère partie

Mathématiques et sciences humaines, tome 118 (1992), p. 33-52

<http://www.numdam.org/item?id=MSH_1992__118__33_0>

Mat. Inf. Sci. hum., (30e année, n° 118, *1992*, p.33-52)

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

1ère partie

Israël-César LERMAN*

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

PRÉAMBULE — *Cet article est divisé en deux parties qui se suivent dans deux numéros consécutifs. La première partie concerne surtout l'aspect conceptuel et la seconde les aspects plus techniques de l'analyse de la forme limite ainsi que les extensions et l'aspect prospectif. En conséquence, les références aux formules dans la deuxième partie, qui commence avec le paragraphe 4, concernent directement la première partie.*

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

This study gives a large synthesis view and prospective on a very general family of association coefficients between descriptive relational variables, that we have elaborated. On the other hand, very accurate technical results are provided. We assume the empirical observation of the descriptive variables on a set O of elementary objects. A given coefficient is obtained by a statistical normalization of a raw association index with respect to a hypothesis of no relation (or independence). The raw index s is conceived from a set theoretic representation of the two relational variables to be compared. The case where the two variables associated are unary, provides a clear setting up of the comparison problem. We particularly analyze the case where the two relations on O, induced by the two descriptive variables to be compared, are binary. The latter case is extremely useful in qualitative data analysis. The normalization of the raw index s takes into account the distribution of the random raw index S under an independence hypothesis. The reduction of the "centred" index [s- E(S) where E denotes the mathematical expectation] is done with the standard deviation $\sqrt{var\,(S)}$. *It is specific expression of the variance*

IRISA, Campus de Beaulieu, 35042 Rennes Cédex.

var(S), which enables to set up the limiting form of an association coefficient, under natural asymptotic conditions. Then, we carefully study the very important cases where the descriptive variables are nominal or ordinal qualitative. The limit expression permits to realize the nature of the normalization, from a purely formal point of view. Next, we take up the study of the general case of the comparison of two q-ary relations. Accurate results are given in the latter context. Finally, we express our current research and their future developement ; more particularly by situating the place of this work in our approach of data analysis by means of hierarchical classification.

PREAMBLE — *This paper is divided into two parts, which will appear consecutively in this and the next issues of the journal. The first part will be mainly concerned with the conceptual framework. The second part will be devoted, on one hand to the analysis of the limit form ; and, on the other hand, to the extension and the prospective of the approach. Then, in the second part -which starts with the section 4-, we refer to the formulas established in part one.*

1. INTRODUCTION GÉNÉRALE ET OBJECTIF DE L'ÉTUDE

Le contexte est celui de la comparaison de deux variables descriptives a et b, observées sur un ensemble O d'objets élémentaires, sur lequel elles sont définies. Généralement, et pour être comparables, a et b se présentent comme deux applications de l'ensemble O dans un ensemble C de codes ou de catégories, qui correspond à ce qu'on appelle, l'échelle des *valeurs*. L'ensemble C peut être muni d'une structure plus ou moins riche ; donnons en quelques exemples :
- préordre total ou partiel sur C
- préordre total ou partiel sur tout ou partie de $C \times C$, ce préordre suppose la possibilité d'appréhender et d'ordonner les différences entre catégories ;
- graphe orienté ou non, sur C ;
- etc...

Pour une vue plus complète, mais dans le cadre de la théorie du mesurage, on pourra consulter [Suppes & Zinnes 1963].

Cette structure sur C induit une relation, éventuellement valuée, sur O et, dans notre démarche, qui s'inscrit dans le cadre de l'Analyse combinatoire des données [Arabie & Hubert, 1992, à paraître], nous ne retenons d'une variable descriptive que la relation qu'elle induit sur O, via la structure dont se trouve muni C. D'autre part et surtout, à cette relation nous associons une partie -éventuellement valuée dans $\mathbb{R}$- de O^q, si q est l'arité de la relation. Ce sous-ensemble de O^q a une structure particulière, compte tenu de la spécificité de la relation. Un exemple simple sur lequel nous reviendrons est celui d'une variable qualitative nominale, où C ne se trouve muni d'aucune structure. Dans ce cas, la variable induit une partition sur l'ensemble O, qui définit une relation d'équivalence à laquelle nous associons son graphe qui est une partie structurée de O^2 ; en effet, n'importe quel sous-ensemble de O^2 ne correspond pas à une relation d'équivalence. Le sous-ensemble de O^2, ainsi associé à la variable qualitative nominale est appelé "représentation *ensembliste* de la variable". Il s'agit bien d'une représentation au sens mathématique du terme ; puisque la correspondance schématisée ci-dessus, détermine bien une bijection entre l'ensemble des variables qualitatives nominales pouvant être définies sur O et l'ensemble des sous-ensembles de O^2 dont chacun peut être interprété comme le graphe d'une relation d'équivalence.

Un autre exemple moins simple -que nous avons effectivement traité [Ouali 1991]- est celui où on suppose que l'ensemble $P_2(C)$ des paires ou parties à deux éléments de C, est muni d'un préordre total qui traduit de façon ordinale les ressemblances entre catégories. Dans ce cas, la représentation ensembliste de la variable est une partie structurée de O^4. Nous allons y revenir.

En vérité, en analyse des données qualitatives, on recouvre la totalité des situations en considérant q = 1,2,3 ou 4. Dans ce dernier cas (q=4), la représentation se situe le plus directement au niveau de $(O \times O)^2$, puisqu'il s'agit -de la manière dont cela se présente dans la pratique - d'une relation binaire sur tout ou une partie de $O \times O$. Une telle relation est généralement de préordre total (ou partiel) et se trouve induite par une variable d'un type nouveau ; la variable "préordonnance" que nous avons [Lerman & Peter (1985), Lerman (1987 a)] en même temps que d'autres [Chah(1984),(1985)] introduite et rendue opérationnelle [Peter(1987), Ouali(1991 a)(1991 b)]. La donnée de ce type de variable suppose une structuration de C au moyen d'un préordre total sur tout ou une partie de $O \times O$. Dans la mesure où on code un tel préordre total à partir d'une fonction ordinale telle que par exemple le "rang moyen", on se trouve ramené à une valuation particulière sur $O \times O$. De sorte que la portée de cet article sera assez générale en se limitant à considérer les cas q=1 et q=2. Nous ferons néanmoins entrevoir des aspects de calcul concernant le cas le plus général [cf. § 6.2.].

Soit la relation discrète ou valuée sur l'ensemble O des objets, définie par une variable descriptive a. Désignons par $R(\alpha)$ sa représentation adoptée au niveau de O^q. Nous avons déjà signalé que dans le premier cas (discret) $R(\alpha)$ est une partie en général structurée de O^q et dans le second cas (valué), $R(\alpha)$ est une pondération ayant en général des caractéristiques particulières, sur O^q. Le cas discret peut, d'un point de vue technique, apparaître comme particulier du cas valué ; il suffit de considérer la valuation définie par la fonction indicatrice de R( ) dans O^q. Cependant, il faut bien distinguer d'un point de vue logique les deux cas. On pourra introduire ici l'ensemble Ω_α de tous les ensembles (resp. valuations) de O^q, susceptibles de représenter une relation (resp. valuation) de même type combinatoire que α.

D'autre part, pour des raisons de clarté combinatoire dans l'expression de la démarche des calculs qui en découlent et de leur interprétation, il importe de donner à la représentation $R(\alpha)$, la forme la plus compacte qui soit. Ainsi, : en reprenant l'exemple ci-dessus, si a est une variable qualitative nominale à k modalités (ou valeurs) respectivement codées 1,2,...j,...,k ; la relation induite par a sur O est une relation d'équivalence ou de partition qu'on peut, pour spécifier, noter . On peut également noter $\pi(O) = \{O_j / 1 \leq j \leq k\}$, la partition sur O définie par π, où $O_j = \{o / o \in O$ et $a(o) = j\}$ est l'ensemble des objets où la valeur de la variable a est j, $1 \leq j \leq k$. Au lieu de $O \times O$, on peut considérer ici l'ensemble plus réduit $O^{\{2\}} = P_2(O)$ des parties à deux éléments de O, pour représenter π par l'ensemble des paires d'objets que la partition réunit ; soit :

$$R(\pi) = \sum \left\{ P_2(0_j) / 1 \leq j \leq k \right\} \qquad (2)$$

(somme ensembliste)

Un autre exemple concerne le cas où a est une variable qualitative totalement ordinale à k modalités qu'on codera respectivement 1,2,...,j,...,k. Il s'agit ici du cas où l'ensemble C ci-dessus des valeurs est totalement ordonné. La relation α induite par a sur O, est une relation de préordre total qu'on peut, pour spécifier, noter ω. Si on admet la représentation de ω par l'ensemble des couples d'objets (o,o') tel que o précède strictement o' pour , on a :

$$R(\pi) = \sum \left\{ O_g \times O_h / 1 \leq g < h \leq k \right\} \qquad (3)$$

Une autre représentation que nous considérerons ci-dessous correspond à une valuation prenant ses valeurs dans {-1,0,1} que nous notons x, définie comme suit :

$$x_{ij} = \begin{cases} 1 & \text{si } o_i < o_j \text{ (pour } \omega) ; \\ 0 & \text{si } o_i \sim o_j \text{ (pour } \omega) ; \\ -1 & \text{si } o_i > o_j \text{ (pour } \omega) ; \end{cases} \qquad (5)$$

où $I = \{1,2,...,i,...,n\}$ indexe l'ensemble O des objets, de cardinal n ; et où, $<$ (resp.$\sim$) indique la relation de précédence stricte (resp. équivalence), relativement à ω.

On peut, toujours dans ce cas de comparaison de deux préordres totaux, proposer, en s'inspirant de [Giakoumakis & Monjardet (1987)], d'autres valuations ; mais, où il importe, pour que notre démarche générale de normalisation statistique s'applique, que le coefficient d'accord entre deux préordres totaux λ, ω et ϖ se présente sous la forme d'un produit scalaire $\lambda<x,y>$, $I \times I$, entre les deux valuations s, x et y, respectivement associées à ω et ϖ.

Pour introduire cette démarche, il y a lieu d'associer à la structure λ sur O - définie par une même variable a - l'ensemble $\mathcal{Q}$ de toutes les structures de même type combinatoire que α. Ainsi, si α est une partition $\pi(O)$ sur O, $\mathcal{Q}$ peut être l'ensemble de toutes les partitions sur O. Il y a lieu d'autre part, de définir une mesure de probabilité $P_{\mathcal{Q}}$ sur $\mathcal{Q}$ de façon à "respecter" les caractéristiques cardinales de la structure α. Ainsi, par exemple, si $\pi(O)$ est une partition dont on suppose - sans restreindre la généralité - qu'elle est en classes étiquetées et si $t(\pi)$ est le type (*i.e.* la suite ordonnée des cardinaux des classes) de la partition ; alors, la mesure de probabilité $P_{\mathcal{Q}}$ peut être concentrée et uniformément répartie sur l'ensemble des partitions en classes étiquetées de même type $t(\pi)$. Nous reviendrons en étant plus précis sur ce point au paragraphe 2 ci-dessous.

Dans ces conditions, le schéma général de *comparaison* (on dit encore d'*association*) de deux variables relationnelles a et b, que nous avons au cours de notre recherche fait émerger, répond au diagramme suivant que nous allons ci-dessous commenter.

$$(a,b) \rightarrow (\alpha,\beta) \rightarrow [R(\alpha), R(\beta)] \in \Omega_\alpha \times \Omega_\beta$$

$$\rightarrow s = s(\alpha,\beta) = card\,[R(\alpha) \cap R(\beta)]$$

$$\text{(resp. } <R(\alpha), R(\beta)>)$$

$$(\alpha,\beta) \xrightarrow{\text{h.a.l.}} (\alpha^*,\beta^*) \in (\mathcal{Q}, P_{\mathcal{Q}}) \times (\mathfrak{B}, P_{\mathfrak{B}})$$

$$\rightarrow S = s(\alpha^*,\beta^*) = card\,[R(\alpha^*) \cap R(\beta^*)]$$

$$\text{(resp. } < R(\alpha^*), R(\beta^*) >)$$

$$Q(\alpha,\beta) = \frac{s - E(S)}{\sqrt{var(S)}} \qquad (5)$$

Figure 1

α, $R(\alpha)$ et Ω_α [resp.β, $R(\beta)$ et Ω_β] ont déjà été définis ci-dessus. $s = s(\alpha,\beta)$ est ce que nous appelons : indice "brut" d'association. Il se présente sous la forme du cardinal d'une intersection $card(X \cap Y)$, respectivement d'un produit scalaire $<X,Y>$, selon que les deux relations de même arité à comparer, sont discrètes et valuées.

$(\mathcal{Q},P_{\mathcal{Q}})$ [resp. $(\mathfrak{B},P_{\mathfrak{B}})$] étant défini et associé à α (resp. β), comme indiqué ci-dessus, α^*(resp. β^*) est un élément aléatoire pris dans $(\mathcal{Q},P_{\mathcal{Q}})$ [resp. $(\mathfrak{B},P_{\mathfrak{B}})$] qui est associé à α(resp. β). D'autre part, α^* et β^* sont indépendants de sorte que nous disons que (α^*, β^*) est associé à (α,β) dans une "hypothèse d'absence de liaison" (h.a.l.).

$S = s(\alpha^*,\beta^*)$ est l'indice brut *aléatoire* dont l'espérance mathématique et la variance sont respectivement notées $E(S)$ et $var(S)$.

$Q(\alpha,\beta)$ est l'indice s "centré et réduit", par rapport à une hypothèse d'indépendance.

On rappellera au paragraphe II ci-dessous les résultats de l'application de ce diagramme dans le cas de la comparaison de deux relations unaires sur O. On considérera en particulier la forme limite de $\mathfrak{Q}(\alpha,\beta)$ dans le cas où l'ensemble O est considéré comme un échantillon aléatoire de taille croissante d'une population infinie $\mathfrak{P}$. Plus précisément, on substituera à l'ensemble O des objets, une suite croissante en taille $\{O^{(n)}/n \geq 1\}$ de parties de $\mathfrak{P}$, telle que la partie $O^{(n)}$, de cardinal n, définit un échantillon aléatoire (exhaustif) de taille n de $\mathfrak{P}$.

L'objet technique de cet article qui reprend le rapport de recherche [Lerman 1987b] consiste à dégager la forme limite de $Q(\alpha^{(n)},\beta^{(n)})$ dans le cas où les deux relations à comparer sont binaires (discrètes ou valuées) ; toutes les deux *symétriques* ou bien, toutes les deux *antisymétriques*. Plus précisément, si $X=\{x_{ij}/(i,j) \in I \times I\}$ et $Y=\{y_{ij}/(i,j) \in I \times I\}$ sont les deux codages ou valuations respectivement définis par $\alpha^{(n)}$ et $\beta^{(n)}$, on suppose l'une ou l'autre des deux situations suivantes :

$$[\forall\ (i,j) \in I \times I]\ (x_{ij} = x_{ji}) \wedge (y_{ij} = y_{ji})\ ; \qquad (6)$$

$$[\forall\ (i,j) \in I \times I]\ (x_{ij} = -\,x_{ji}) \wedge (y_{ij} = -\,y_{ji}) \qquad (7)$$

Signalons que l'indice centré $[s - E(S)]$ donne exactement le numérateur du coefficient τ de M.G. Kendall [Kendall 1970], en cas de comparaison de deux variables "rang" a et b qui induisent deux ordres totaux et stricts sur l'ensemble O des objets. La normalisation proposée par M.G. Kendall consiste à rapporter l'indice centré à sa valeur maximale. Certains chercheurs tels que L. Hubert [Hubert 1983] considèrent qu'un coefficient d'association entre deux variables qualitatives doit nécessairement avoir la même forme que celle proposée par M.G. Kendall. C'est-à-dire, dans le contexte général décrit ci-dessus :

$$\tau(\alpha,\beta) = \frac{s(\alpha,\beta) - E\,[s(\alpha^*,\beta^*)]}{\mathfrak{M}\,[s(\alpha^*,\beta^*) - E\,[s(\alpha^*,\beta^*)]}\ , \qquad (8)$$

où $\mathfrak{M}[s(\alpha^*,\beta^*)]$ est la valeur extrême maximale de la distribution de $s(\alpha^*,\beta^*)$.

Nous avons bien considéré, dans des situations parfois très difficiles, le problème de la découverte de $\mathfrak{M}[s(\alpha^*,\beta^*)]$ [Lerman 1987b, 1988], en y apportant une solution de type algorithmique récursive. Les deux cas de figure que nous avons étudiés sont d'une part, celui où α et β sont des partitions et d'autre part, celui où et sont des préordres totaux. Le premier cas est repris par H. Messatfa dans sa thèse [Messatfa 1990] avec un apport méthodologique en termes de programmation linéaire. Le second cas nourrit l'article de V. Giakoumakis et B. Monjardet [Giakoumakis & B. Monjardet 1987]. Ces auteurs développent un point de vue métrique unificateur englobant un large ensemble de coefficients proposés ; mais où, la normalisation - au dénominateur du coefficient - par une quantité majorant la valeur maximale du numérateur, ne tient pas le plus étroitement compte de la contrainte de structure que représente un préordre total.

Pour nous, les deux types de normalisation [cf. (5) et (8) ci-dessus], méritent d'être considérés et étudiés. Nous serons ici concernés par la normalisation statistique au moyen de $\sqrt{var\,(S)}$, pour la définition d'un coefficient d'association.

Ainsi, par rapport au diagramme précédent (cf. Figure 1), il est intéressant de situer les différents coefficients d'accord proposés dans la littérature. On se posera notamment la question

de connaître leur espérance mathématique et leur variance par rapport à une forme donnée de l'hypothèse d'absence de liaison.

Quelle que soit l'expression formelle proposée d'un coefficient d'association entre variables qualitatives, Goodman et Kruskal [Goodman & Kruskal 1963] proposent de façon intéressante l'étude de la distribution d'échantillonnage de l'indice calculé au niveau de l'ensemble O des objets, par rapport à sa valeur théorique au niveau d'une population $\mathfrak{P}$, dont O est considéré comme un échantillon aléatoire. Plus précisément, on considère une suite infinie $\{O_j^{(n)} / j \geq 1\}$ d'échantillons aléatoires de même taille n, d'une population de taille infinie $\mathfrak{P}$. Si $\chi_j(\alpha,\beta)$ désigne la valeur calculée d'un coefficient d'association entre les deux structures et sur $O_j^{(n)}$, respectivement définies par les deux variables a et b à comparer, il y a lieu d'étudier la distribution $\{\chi_j(\alpha,\beta) / j \geq 1\}$ par rapport à la valeur $\chi(\alpha,\beta)$ définie au niveau de la comparaison entre α et β sur $\mathfrak{P}$. Curieusement, les auteurs précités n'avaient pas considéré une telle étude dans le cas de la comparaison de variables (on dit encore "attributs") définissant des relations unaires sur l'ensemble décrit. Nous avons montré dans [Lerman 84] toute la richesse de cette étude que nous faisons intervenir après la découverte de l'expression formelle du coefficient $Q(\alpha,\beta)$, à partir du diagramme de la figure 1 ci-dessus ; et, précisément, de la mise en évidence de la forme limite de $Q[\alpha^{(n)}, \beta^{(n)}]$ par rapport à n dans des conditions exprimées précédemment (cf. §2 ci-dessous). D'où, un des aspects de l'intérêt de dégager la forme limite de $Q[\alpha^{(n)}, \beta^{(n)}]$ dans le cas où α et β sont des relations binaires de différents types.

Un autre aspect statistique important plus directement lié à la nature de notre approche a un caractère plus intrinsèque. Il concerne, pour n fixé, l'étude de la distribution de l'indice aléatoire $Q(\alpha^*,\beta^*)$:

$$Q(\alpha^*,\beta^*) = \frac{s(\alpha^*,\beta^*) - E(S)}{\sqrt{var(S)}} \quad ; \qquad (9)$$

et, de la tendance asymptotique d'une telle distribution lorsque le cardinal n augmente. Dans des conditions assez générales et le plus fréquemment -mais pas toujours- quant à la nature combinatoire et cardinale des deux structures à associer, une telle distribution est asymptotiquement normale $\mathfrak{N}(0,1)$ [Wald et Wolfowitz 1944, Noether 1949, Hajek 1961, Lerman 1977, 1981, Mielke 1979, Daudé 1990].

2. COMPARAISON DE DEUX VARIABLES RELATIONNELLES UNAIRES

2.1. Structure aléatoire

Il s'agit du cas où les deux variables (on dit également "attributs") a et b sont toutes les deux de "présence-absence" (on dit encore booléennes) ou bien, "quantitatives". Dans le premier cas (a,b) induit un couple de parties $[O(a),O(b)]$ de parties de l'ensemble O des objets et dans le second cas (a,b) induit un couple de valuations sur O, que nous noterons (x_I,y_I) ; où

$$x_I=\{x_i / i \in I\}$$
$$(\text{resp. } y_I = \{y_i / i \in I\}$$

où, rappelons-le, $I = \{1,2,...,i,...,n\}$ indexe l'ensemble O.

Relativement à une même variable observée définissant une structure η sur O (partie ou valuation), nous allons spécifier la correspondance

$$\eta \rightarrow \eta^* \in (E,P_E)$$

considérée au paragraphe 1. η^* est la structure aléatoire associée à η de façon à "respecter" la cardinalité de η. Cette structure est ici soit une partie d'un ensemble, soit une valuation sur ce dernier.

(i) Partie aléatoire
Désignons par $\hat{\mathcal{E}}$ le sous-ensemble de O représentant η. On pose $m = card(\hat{\mathcal{E}})$. Au couple $(\hat{\mathcal{E}},O)$ on associe - comme cela sera précisé ci-dessous - un couple aléatoire d'ensembles (X,O^*), où $X \subset O^*$.

Nous avons dégagé trois modèles aléatoires fondamentaux de choix du couple (X,O^*) chacun respectant d'une certaine façon la cardinalité (m,n) du couple $(\hat{\mathcal{E}},O)$ [Lerman, Gras et Rostam 1981, Lerman 1981]. Ces trois modèles sont respectivement notés N_1, N_2 et N_3 ; N_1 (resp. N_2) est plus "diffus" que N_2 (resp. N_3). Nous allons rapidement les rappeler.

N_1 : Pour ce modèle aléatoire, la réalisation O_0 de O^* est l'ensemble O des objets. X est un élément aléatoire dans l'ensemble $P_m(O)$ des parties de O de même cardinal m. Il y a $\binom{n}{m}$ de telles parties. D'autre part, la probabilité est uniformément répartie sur $P_m(O)$.

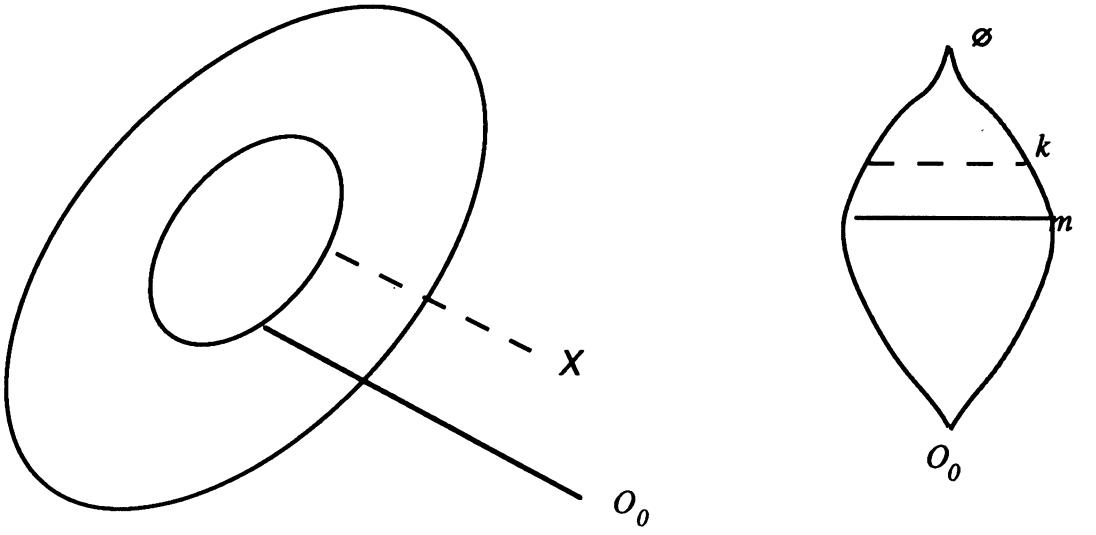

Figure 2

N_2 : Pour ce modèle aléatoire la réalisation O_0 de O^* est toujours l'ensemble O des objets. Si on regarde l'ensemble 2^O des parties de O organisé par inclusion, où un même niveau est formé des sous-ensembles de même cardinal, le modèle N_1 correspond à charger toute la probabilité sur le niveau m du treillis ; alors que le modèle N_2 diffuse la probabilité sur l'ensemble des niveaux. Le niveau k est affecté de la probabilité binomiale :

$$\binom{n}{m}\mu^k(1-\mu)^{n-k} \quad \text{où} \quad \mu = m\ /\ n, 0 \le k \le n\,. \qquad (1)$$

D'autre part, pour k fixé, cette dernière probabilité est uniformément répartie sur l'ensemble des $\binom{n}{k}$ sommets du niveau k. Ainsi, tout se passe comme s'il s'agit d'un modèle aléatoire de choix à deux pas ; le premier concerne le choix du niveau et le second, celui de l'élément (qui est une partie de cardinal k de O) dans ce niveau.

N_3: Il s'agit ici d'un modèle aléatoire à trois pas :

- Le premier consiste à associer à O un ensemble aléatoire O^* dont on spécifie seulement la loi de la variable aléatoire $\mu = card(O^*)$ qui est de Poisson de paramètre n :

$$Pr\{v=l\} = \frac{n^l}{l!} e^{-n} \qquad (2)$$

- Sachant $O^* = O_0$ de cardinal l_0, le deuxième pas consiste dans le choix aléatoire d'un niveau associé à $\mathfrak{G}$ dans l'ensemble des parties de O_0. Ce choix se fait selon le modèle binomial

$$Pr\{K = k\} = \binom{l_0}{k} \mu^k \ (1 - \mu)^{l_0 - k} \qquad (3) :$$

où, rappelons-le, $\mu = m/n$ et où $k = 1,2,...,l_0$.

- Sachant $O^* = O_0$ de cardinal l_0 et $K = k$, le choix aléatoire de X se fait uniformément au hasard sur le niveau k du treillis des parties de O_0.

(ii) Valuation aléatoire

Nous allons à présent mettre en évidence les correspondants des modèles N_1, N_2 et N_3 dans le cas où la donnée est une valuation :

$$x_I = \{x_i / i \in I\}, \qquad (4)$$

où, rappelons-le encore une fois, $I = \{1,2,...,i,...,n\}$ indexe l'ensemble O des objets.

N_1: Pour ce modèle, l'élément aléatoire de base sera une permutation σ sur I ; c'est-à-dire, une autobijection sur I. σ est un élément pris dans l'ensemble G_n, muni d'une probabilité uniformément répartie, de toutes les permutations sur I. Il y a $n!$ permutations $[n! = card(G_n)]$.

L'image :

$$\sigma(I) = [\sigma(1), \sigma(2),..., \sigma(i),...,\sigma(n)] \qquad (5)$$

est également appelée "permutation" de I.

A la suite $(x_1,x_2,...,x_i,...x_n)$ des valeurs observées, σ associe la suite aléatoire :

$$x_{\sigma(I)} = [x_{\sigma(1)}, x_{\sigma(2)},..., x_{\sigma(i)},..., x_{\sigma(n)}] \qquad (6)$$

N_2: Pour ce modèle, à la suite $(1,2,...,i,...,n)$ des étiquettes, on associe une suite aléatoire de n étiquettes que nous noterons comme suit :

$$I^* = (i_1^*, i_2^*,..., i_j^*,..., i_n^*) \qquad (7)$$

où les i_j^*, $1 \le j \le n$, sont mutuellement indépendants et où pour chaque j, $1 \le j \le n$, on a :

$$Pr(i_j^* = k) = \frac{1}{n} , \qquad (8)$$

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

L'objet aléatoire (7) permet d'associer à la suite des valeurs observées $(x_1,x_2,...,x_i,...,x_n)$, la suite aléatoire :

$$x_{I^*} = (x_{i_1^*}, x_{i_2^*}..., x_{i_j^*},..., x_{i_n^*}) \qquad (9)$$

N_3: Pour ce modèle, on commence par associer à n, un entier aléatoire ν dont la loi est de Poisson de paramètre n [cf.(2) ci-dessus].

Pour $\nu = l$ fixé, on considère un vecteur aléatoire aux composantes mutuellement indépendantes :

$$I_l^* = (i_1^*, i_2^*, \ldots, i_j^*, \ldots, i_l^*) \qquad (10)$$

qu'on peut supposer déterminé après l tirages aléatoires indépendants ; où, pour tout $j = 1,2,\ldots,l$;

$$Pr\ (i_j^* = i) = \frac{1}{n}\ , \qquad (11)$$

quel que soit $i = 1,2,\ldots,n$.

Le troisième pas du modèle est semblable à celui N_2 ci-dessus. Plus précisément, pour l fixé, à la suite I_l^* des étiquettes aléatoires, on associe la suite

$$x_{I_l^*} = (x_{i_1^*}, x_{i_2^*} \ldots, x_{i_j^*}, \ldots, x_{i_n^*}) \qquad (12)$$

de l variables aléatoires indépendantes et de même loi définie par la distribution empirique $\{x_i / i \in I\}$.

Finalement, à la suite $(x_1, x_2, \ldots, x_i, \ldots, x_n)$ des valeurs observées se trouve associée la suite

$$x_{I_\nu^*} = (x_{i_1^*}, x_{i_2^*} \ldots, x_{i_j^*}, \ldots, x_{i_\nu^*}) \qquad (13)$$

doublement aléatoire en raison de l'aléa sur et de celui (11) ci-dessus.

Revenons à (*i*) ci-dessus. On peut montrer, qu'en représentant le sous-ensemble $\mathfrak{S}$ de O par une valuation, au moyen d'un vecteur (formé de zéros et de uns) indicateur de $\mathfrak{S}$ dans O, on peut retrouver, à partir de N_1, N_2 et N_3 de (*ii*), N_1, N_2 et N_3 de (*i*). Mais il importe de distinguer clairement (*i*) de (*ii*) ; faute de quoi, tous les phénomènes combinatoires inhérents à (*i*), seront marqués dans (*ii*).

Le modèle aléatoire de choix correspondant à N_2 rappelle la technique de rééchantillonnage du "Bootstrap" [Efron 1986]. Cependant nous y parvenons de façon tout à fait indépendante à partir de [Lerman, Gras & Rostam 1981_a Lerman 1981_b], dans une vision, un contexte et un objectif fondamentalement différents.

2.2. Coefficients d'association

Nous allons considérer les coefficients d'association $Q(\alpha,\beta)$ qu'on obtient dans le cadre du schéma de la figure 1 du paragraphe 1 [cf. (5) § 1]. Nous supposons que α^* et β^* sont produits selon le même type N_i de modèle aléatoire ($i = 1,2$ ou 3). H_i désignera l'hypothèse d'absence de liaison (on dit encore d'indépendance) correspondante au modèle aléatoire de choix N_i ($i = 1$, 2 ou 3).

(i) Cas discret

Les deux attributs a et b à associer sont ici de "présence-absence" (on dit encore booléens). La représentation de l'attribut a (resp. b) est le sous-ensemble $O(a)$ [resp. $O(b)$] des objets où a (resp. b) est présent.

Désignons ici par $n(a)$ [resp. $n(b)$] le cardinal de l'ensemble $O(a)$[resp. $O(b)$] et par $p(a) = n(a)/n$ [resp. $p(b) = n(b)/n$] la proportion des objets où l'attribut a (resp. b) est présent. Introduisons de plus l'attribut $\bar{a}$ (resp. $\bar{b}$) opposé à a (resp. b). $O(\bar{a})$ [resp.$O(\bar{b})$] est le sous-ensemble des objets complémentaire de $O(a)$ [resp. $O(b)$] dans l'ensemble O. Nous noterons $n(\bar{a})$ [resp. $n(\bar{b})$] le cardinal de $O(\bar{a})$ [resp.$O(\bar{b})$] et par $p(\bar{a}) = n(\bar{a})/n$ [resp.$p(\bar{b}) = n(\bar{b})/n$] la proportion des objets où l'attribut a (resp. b) est absent.

L'indice brut se met ici sous la forme :

$$s = s(\alpha,\beta) = card[O(a) \cap O(b)] \qquad (14)$$

et, par conséquent, celui aléatoire, sous la forme :

$$S = s(\alpha^*,\beta^*) = card[X \cap Y] \qquad (15)$$

où (X,Y) est un couple de parties aléatoires indépendantes, associé au couple $[O(a),O(b)]$ dans l'hypothèse d'absence de liaison H_i (i = 1,2,3) (*cf.* ci-dessus).

Nous avons établi [Lerman, Gras et Rostam 1981, Lerman 1981] que la loi de probabilité de S est hypergéométrique, binomiale ou de Poisson, selon que i =1,2 ou 3. Désignons par

$$E_i [s(\alpha^*,\beta^*)] = E[s(\alpha^*,\beta^*) / H_i] \qquad (16)$$

et

$$var_i [s(\alpha^*,\beta^*)] = var[s(\alpha^*,\beta^*) / H_i] , \qquad (17)$$

On a, quel que soit i,

$$E_i [s(\alpha^*,\beta^*)] = \frac{n(a)\, n(b)}{n} = np(a)p(b), \qquad (18)$$

mais, $var_i[s(\alpha^*, \beta^*)]$ dépend de i :

$$var_1 [s(\alpha^*,\beta^*)] = \frac{n(a)\, n(\bar{a})n(b)\, n(\bar{b})}{n^2(n-1)}$$

$$= \frac{n^2}{(n-1)} p(a)\, p(\bar{a})p(b)\, p(\bar{b}) , \qquad (19)$$

$$var_2 (s (\alpha^*,\beta^*) = n\, p(a)\, p(b)\, [1 - p(a)\, p(b)] , \qquad (20)$$

et

$$var_3 (s (\alpha^*,\beta^*) = n\, p(a)\, p(b), \qquad (21)$$

En confondant $(n-1)$ et n, l'indice "centré et réduit" (statistiquement) se met dans chacun des cas sous la forme

$$Q_i(a,b) = \sqrt{n}\ r_i(a,b), \quad i = 1, 2 \text{ ou } 3 \qquad (22)$$

où

$$r_1(a,b) = \frac{p(a \wedge b) - p(a)p(b)}{\sqrt{p(a)\, p(\bar{a})p(b)\, p(\bar{b})}} \qquad (23)$$

$$= \frac{p(a \wedge b)\, p(\bar{a} \wedge \bar{b}) - p(b \wedge \bar{b})p(\bar{a} \wedge b)}{\sqrt{p(a)\, p(\bar{a})p(b)\, p(\bar{b})}}$$

avec des notations que l'on comprend ; c'est-à-dire, où $e = c \wedge d$ est l'attribut booléen résultant de la conjonction des deux attributs booléens c et d.

Le coefficient $r_1(a,b)$ n'est autre que celui de K. Pearson. On a d'autre part,

$$r_2(a,b) = \frac{p(a \wedge b) - p(a)\,p(b)}{\sqrt{p(a)\,p(b)\,[1 - p(a)\,p(b)]}} \tag{24}$$

et

$$r_3(a,b) = \frac{p(a \wedge b) - p(a)\,p(b)}{\sqrt{p(a)\,p(b)}} \tag{25}$$

On peut aisément établir que

$$p(a \wedge b) - p(a)\,p(b) \leq \sqrt{p(a)\,p(\overline{a})\,p(b)\,p(\overline{b})}\,; \tag{26}$$

D'autre part, on a

$$p(a)\,p(\overline{a})\,p(b)\,p(\overline{b}) \leq p(a)\,p(b)\,[1 - p(a)\,p(b)] \leq p(a)\,p(b) \tag{27}$$

de sorte que

$$|r_1(a,b)| \geq |r_2(a,b)| \geq |r_3(a,b)|\,; \tag{28}$$

toutes les inégalités ci-dessus étant en général strictes.

On peut remarquer en passant que si

$$p(a) < p(\overline{a}) \text{ et } p(b) < p(\overline{b})$$

alors :

$$r_1(\overline{a}\,,\,\overline{b}\,) = r_1(a,b) \tag{29}$$

mais

$$r_3(\overline{a}\,,\,\overline{b}\,) < r_3(a,b) \tag{30}$$

Ainsi, toutes choses logiquement égales par ailleurs, r_3 quantifie davantage la ressemblance entre attributs rares. Cette propriété traduit notre perception intuitive de la ressemblance. Elle se trouve en accord avec l'optique de la théorie de l'information.

(ii) Cas valué

Les deux attributs v et w à associer sont ici "quantitatifs" ou "numériques". Conformément à ci-dessus, nous désignons par (x_I, y_I) le couple de valuations sur l'ensemble O des objets, défini par le couple de variables :

$$x_I = \{x_I / \in I\} \text{ et } y_I / i \in I\}, \tag{31}$$

où $I = \{1,2,...,i,...,n\}$ indexe O.

Désignons par μ_x et σ_x^2 (resp. μ_y et σ_y^2 la moyenne et la variance de la distribution des valeurs x_I(resp. y_I). Soient à présent

$$\left.\begin{aligned} x_J^* &= (x_1^*, x_2^*, \ldots, x_j^*, \ldots, x_v^*) \\ \text{et} \\ y_J^* &= (y_1^*, y_2^*, \ldots, y_j^*, \ldots, y_v^*) \end{aligned}\right\} \tag{32}$$

les deux valuations aléatoires obtenues dans le cadre de l'hypothèse d'absence de liaison H_i (i = 1,2 ou 3).

Si l'indice brut se met sous la forme :

$$s = \langle x_I, y_I \rangle = \sum_{1 \le i \le n} x_i y_i \; ; \qquad (33)$$

celui, aléatoire, s'écrit :

$$S = \langle x_J^*, y_J^* \rangle = \sum_{1 \le j \le v} x_j^* y_j^* \; ; \qquad (34)$$

On pourra aisément préciser sa forme dans le cadre de chacun des modèles aléatoires N_1,N_2 et N_3 [cf. (*ii*), § 2.1], respectivement sous-jacents à H_1, H_2 et H_3.

Ainsi, sous l'hypothèse H_1, $v = n$ et S s'écrit :

$$s(\sigma,\tau) = \sum_{1 \le i \le n} x_{\sigma(i)} x_{\tau(i)} \qquad (35)$$

où σ et τ sont deux permutations aléatoires indépendantes sur I, conformes au modèle N_1. En vérité, la distribution de S est la même que celle, de l'une ou de l'autre des deux variables aléatoires duales :

$$s(I_d,\tau) = \sum_{1 \le i \le n} x_i \, y_{\tau(i)} \quad \text{et} \quad s(\sigma,I_d) = \sum_{1 \le i \le n} x_{\sigma(i)} y_i \qquad (36)$$

L'étude de la tendance asymptotique d'une telle distribution fait l'objet d'un célèbre théorème de la statistique non paramétrique [Wald & Wolfowitz 1944], [Noether 1949] et [Hajek 1961]. C'est, sous des conditions très générales que cette tendance asymptotique est normale. L'espérance mathématique et la variance exactes de la distribution commune de $s(\sigma,\tau)$,$s(I_d,, \tau)$ et $s(\sigma, I_d)$ sont :

$$E(S) = n \, \mu_x \, \mu_y \qquad (37)$$

et

$$var_1(S) = \frac{n^2}{n-1} \, \sigma_x^2 \, \sigma_y^2, \qquad (38)$$

où on a noté S la variable aléatoire commune.

L'indice brut s, centré et réduit se met sous la forme [cf. (5) § 1] :

$$Q_1(v,w) = \frac{s - E(S)}{\sqrt{var_1(S)}} = \sqrt{n-1} \; r_1(v,w) \qquad (39)$$

où $r_1(v,w)$ n'est autre que le coefficient de corrélation de Bravais-Pearson.

Revenons à (*i*) ci-dessus, si on code le sous-ensemble $O(a)$ [resp.$O(b)$] au moyen d'une valuation v (resp. w) qui correspond à sa fonction indicatrice dans O, il est clair qu'on retrouve $r_1(a,b)$ [cf. (23)] à travers $r_1(v,w)$. On remarquera toutefois que lorsque la permutation σ (resp. τ) décrit l'ensemble G_n des n! permutations sur $I = \{1,2,...,i,...,n\}$, une même partie X (resp. Y) associée à $O(a)$ [resp. $O(b)$] est décrite $[n(a)!]x[n(\bar{a})!]$ (resp. $[n(b)!] \times [n(\bar{b})!]$) fois.

Considérons à présent l'hypothèse H_2 d'absence de liaison. On a toujours $\nu = n$ et ; conformément à (*ii*) du paragraphe 2.1, l'indice aléatoire S se met sous la forme

$$s(I^*,I'^*) = \langle x_{I^*}, y_{I'^*} \rangle = \sum_{1 \le j \le n} x_{i_j^*} \, y_{i'^*_j} \qquad (40)$$

où $I^* = (i_1^*, i_2^*, \ldots, i_j^*, \ldots, i_n^*)$ et $I'^* = (i'^*_1, i'^*_2, \ldots, i'^*_j, \ldots, i'^*_n)$ sont deux suites aléatoires indépendantes obtenues conformément au modèle N_2[cf.(*ii*) § 2.1].

En utilisant le théorème central limite, on peut montrer que sous des conditions très générales, la tendance asymptotique de la loi de S est la loi normale. La moyenne et la variance exactes de S sont, respectivement,

$$E(S) = n \, \mu_x \, \mu_y \qquad (41)$$

et

$$var_2(S) = n(\sigma_x^2 \, \sigma_y^2 + \mu_y^2 \, \sigma_y^2 + \mu_y^2 \, \sigma_x^2) \qquad (42)$$

On pourra noter avec intérêt que $E(S)$ et var(S) se réduisent respectivement à $n \, p(a)p(b)$ et à $n \, p(a)p(b) \, [1\text{-}p(a)p(b)]$ [cf. (18) et (20) de (*i*) ci-dessus], dans le cas où les valuations v et w correspondent respectivement aux fonctions indicatrices de $O(a)$ et de $O(b)$. Ici encore, on pourra comparer les cardinalités respectives de l'espace de l'échantillon dans les cas (*i*) et (*ii*). Il s'agit de 2^n pour le choix de la partie aléatoire X (resp. Y) associée à $O(a)$ [resp. $O(b)$] ; mais il s'agit de n^n pour le choix de x_{I^*} (resp. $y_{I'^*}$).

Il reste à considérer l'hypothèse d'absence de liaison H_3, où v devient une variable aléatoire entière de Poisson de paramètre n [cf. (2), (*i*), § 2.1]. Dans ces conditions, l'indice brut aléatoire S se met sous la forme

$$s(I_\nu^*, I_\nu'^*) = \sum_{1 \le j \le \nu} x_{i_j^*} \, y_{i'^*_j} \qquad (43)$$

Ici encore, on pourra établir aisément que la loi de probabilité de S tend -dans des conditions très générales- vers la loi normale. D'autre part, un calcul simple montre que l'espérance mathématique et la variance de S se mettent respectivement sous la forme :

$$E(S) = n \; \mu_x \, \mu_y \qquad (44)$$

et

$$var_3(S) = n(\sigma_x^2 + \mu_x^2)(\sigma_y^2 + \mu_y^2) \qquad (45)$$

On peut alors constater que $E(S)$ et $var(S)$ se réduisent respectivement tous les deux à $n \, p(a)p(b)$ [cf. (18) et (21) de (*i*) ci-dessus], dans le cas où les valuations v et w correspondent respectivement aux fonctions indicatrices de $O(a)$ et de $O(b)$.

Considérons à présent les indices statistiques $Q_2(v,w)$ et $Q_3(v,w)$, résultant du centrage et de la réduction de l'indice brut s, respectivement par rapport aux hypothèses H_2 et H_3(cf. (5) § 1]. On obtient les formes suivantes :

$$Q_2(v,w) = \sqrt{n} \, r_2(v,w) \qquad (46)$$

et

$$Q_3(v,w) = \sqrt{n} \, r_3(v,w) \qquad (47)$$

où

$$r_2(v,w) = \frac{p(v,w) - \mu_x \, \mu_y}{\sqrt{\sigma_x^2\sigma_y^2 + \mu_x^2\sigma_y^2 + \mu_y^2\sigma_x^2}} \qquad (48)$$

et

$$r_3(v,w) = \frac{p(v,w) - \mu_x \mu_y}{\sqrt{(\sigma_x^2 + \mu_x^2)(\sigma_y^2 + \mu_y^2)}} \qquad (49)$$

où $p(v,w)$ représente le rapport s/n ; en d'autres termes, le numérateur commun de $r_2(v,w)$ et de $r_3(v,w)$ représente la covariance entre les deux variables numériques v et w.

Comme $r_1(a,b)$,$r_2(a,b)$ et $r_3(a,b)$ [cf. (23), (24) et (25)], $r_1(v,w)$, $r_2(v,w)$ et $r_3(v,w)$ [cf.(39),(48) et (49)] sont des coefficients "purs" qui ont le sens d'une corrélation. Plus précisément et d'autre part, par rapport à l'optique considérée dans l'introduction, d'un échantillon aléatoire O de taille croissante, d'une population de taille infinie $\mathfrak{P}$, $r_1(a,b)$, $r_2(a,b)$ et $r_3(a,b)$ [resp. $r_1(v,w)$, $r_2(v,w)$ et $r_3(v,w)$] tendent vers leurs valeurs théoriques, $\rho_1(a,b)$, $\rho_2(a,b)$ et $\rho_3(a,b)$ [resp. $\rho_1(v,w)$, $\rho_2(v,w)$ et $\rho_3(v,w)$] calculées au niveau de $\mathfrak{P}$.

3. COMPARAISON DE DEUX VARIABLES RELATIONNELLES BINAIRES.

3.1. Structure aléatoire

Faisant suite à l'introduction générale, une variable relationnelle binaire sur l'ensemble O des objets élémentaires, se trouve représentée par une partie structurée ou une valuation plus ou moins spécifique de $O \times O$. En continuant à désigner par $I = \{1,2,...,i,...,n\}$ l'ensemble d'indexation de O, on peut exprimer tout de suite que la structure aléatoire binaire sur O, va être le reflet de celle définie à partir de I, au moyen des modèles N_1, N_2 ou N_3 [cf. (*ii*), §2.1 ci-dessus].

Nous allons toutefois et plus précisément, directement définir la structure aléatoire binaire, en distinguant comme ci-dessus, le cas discret du cas valué.

(*i*) *Partition aléatoire*
Sans restreindre la généralité, nous supposerons que la partition donnée sur O est en classes étiquetées :

$$\pi = \{O_j \,/\, 1 \leq j \leq k\} \qquad (1)$$

Nous désignons par $t(\pi)=(n_1,n_2,...,n_i,...,n_k)$ le *type* de la partition π ; c'est-à-dire, la suite des cardinaux de ses classes :

$$n_j = card(O_j), \qquad (2)$$

pour tout $j = 1,2,...,k$.

N_1 : le modèle aléatoire N_1 consiste à associer à π, une partition aléatoire π_1^* dans l'ensemble $P(n;t)$ des partitions en classes étiquetées de type $t = t(\pi)$; on a

$$card\ [P(n\ ;t)] = \frac{n!}{n_j!...n_j!...n_{k!}}. \qquad (3)$$

Nous avons envisagé dans [Lerman 1981] l'extension des modèles aléatoires N_2 et N_3 pour le choix d'une partition aléatoire associée à π. F. Daudé [Daudé 1990] a précisé une telle extension au niveau de la construction d'une table de contingence aléatoire, croisant deux variables qualitatives. Notre expression ci-dessous sera assez différente ; en effet, elle s'ajustera à l'expression ensembliste qui précède et concernera de façon séparée et indépendante chacun des deux côtés du tableau de contingence.

N_2 : On commence par considérer la suite des proportions $\{p_j/1 \leq j \leq k\}$, où $p_j = n_j/n$. On considère ensuite les probabilités multinomiales de la forme :

$$\frac{n!}{m_1!m_2!...m_{kj}!...m_k!}\, p_1^{m_1} p_2^{m_2} \times ... \times p_j^{m_j} \times ... \times p_k^{m_k}. \quad (4)$$

Cette probabilité qu'on pourra noter $P(m_1,m_2,...,m_j,...,m_k)$ est celle que la partition aléatoire

$$\pi_2^* = \{O_j^*/1 \leq j \leq k\} \quad (5)$$

associée à π, soit de type $(m_1,m_2,...,m_j,...,m_k)$; c'est-à-dire, que

$$(\forall j = 1,...,k)\ ,\ card(O_j^*) = m_j \quad (6)$$

Maintenant, étant donnée la suite $(m_1,m_2,...,m_j,...,m_k)$, la partie O_1^* est un élément aléatoire pris uniformément au hasard dans l'ensemble des parties, de même cardinal m_1, de O. Pour $O_1^* = O_1^0$, O_2^* est un élément aléatoire dans l'ensemble, muni d'une probabilité uniforme, des parties, de même cardinal m_2, de l'ensemble $(O - O_1^0)$. Plus généralement, pour

$$O_1^* = O_1^0\ ,\ O_2^* = O_2^0,...,O_{j-1}^* = O_{j-1}^0$$

O_j^* est un élément aléatoire dans l'ensemble, muni d'une probabilité uniforme, des parties, de même cardinal m_j, de l'ensemble

$$O - \left(\sum_{1 \leq h \leq j-1} O_h^0 \right), \quad 1 \leq j \leq k \qquad \text{(somme ensembliste)}$$

N_3 : Le premier pas de ce modèle aléatoire à trois pas a déjà été exprimé en (*i*) du paragraphe 2.1.

- Sachant $O^* = O_0$ de cardinal l_0, les deux pas suivants du modèle sont ceux de N_2, avec, au lieu de n, on a l_0 ; les valeurs des $p_j = n_j/n$, $1 \leq j \leq k$, restant les mêmes.

(ii) Valuation aléatoire

Introduisons ici l'ensemble

$$I^{[2]} = \{\ (i,j)\ /\ 1 \leq i \neq j \leq n\} \quad (7)$$

des couples d'indices à composantes distinctes. Une valuation binaire qui correspond à un graphe orienté valué sur l'ensemble O des objets, se présentera pour nous soit sous la forme d'une pondération x sur $I^{[2]}$:

$$\left\{x_{ij}/(i,j) \in I^{[2]}\right\}, \quad (8)$$

soit sous la forme d'une pondération x sur $I^2 = I \times I$:

$$\left\{x_{ij}/(i,j) \in I \times I\right\}, \quad (9)$$

mais, dans ce dernier cas, on supposera que la valeur de x sur la diagonale $\{(i,i)/i \in I\}$ est constante ($x_{ii} = c$, pour tout i =1,2,...,n). L'indice final [cf. (5) § 1] est invariant quelle que soit la valeur finie de la constante. Néanmoins, il faut être cohérent ; et, dans la situation qui nous concernera ici de la comparaison de deux valuations symétriques ou antisymétriques [cf. (6) et (7), §1], nous poserons :

$(\forall\ i \in I)\ x_{ii} = 1$, dans le cas symétrique
(resp. $x_{ii} = 0$ dans le cas antisymétrique).

On peut dès lors signaler que la forme (8) est plus adaptée dans l'introduction de l'hypothèse d'absence de liaison à caractère permutationnel N_1 ; alors que (9) rend plus souple l'introduction des hypothèses d'absence de liaison N_2 et N_3.

Comme nous l'avons mentionné ci-dessus, N_1, N_2 et N_3 qui s'expriment au niveau de I, ont déjà été spécifiés ci-dessus [cf. (*ii*) § 2.1].

La valuation aléatoire associée à (8) dans le cadre de N_1 s'exprime comme suit :

$$\left\{x_{\sigma(i)\sigma(j)} / (i,j) \in I^{[2]}\right\} \qquad (10)$$

où σ est une permutation aléatoire sur I qui a déjà été spécifiée en (*ii*) du paragraphe 2.1.

La valuation aléatoire associée à (9) dans le cadre de N_2 prend la forme suivante :

$$\{x_{i_j^* i_k^*} / 1 \leq j,k \leq n\} \qquad (11)$$

où $I^* = (i_j^* / 1 \leq j \leq n)$ est une suite de n indices aléatoires indépendants déjà définie ci-dessus [cf. (7) de (*ii*), § 2.1].

La valuation aléatoire associée à (9) dans le cadre du modèle N_3 s'écrit comme suit :

$$\{x_{i_j^* i_k^*} / 1 \leq j,k \leq v\} \qquad (12)$$

où la suite aléatoire $I^*_v = (i^*_j / 1 \leq j \leq v)$ est précisée dans (*ii*) du paragraphe 2.1.

En représentant la partition π au moyen d'une application f, faisant correspondre à chaque objet, l'étiquette de la classe à laquelle il appartient, on retrouve les trois modèles aléatoires N_1, N_2 et N_3 du cadre (*i*) ci-dessus, à partir de celui (*ii*) où on se trouve.

3.2. Indice brut aléatoire et suite à donner

Nous allons considérer le cas de la comparaison d'un couple de valuations sur $I^{[2]}$(resp.IxI) de la forme (8) [resp.(9)]. Nous désignerons ce couple de valuations comme suit :

$$X = \left\{x_{ij} / (i,j) \in J\right\} \qquad (13)$$

et

$$Y = \left\{y_{ij} / (i,j) \in K\right\} \qquad (14)$$

où K désigne $I^{[2]}$ ou bien $I \times I$. D'ailleurs, comme nous l'avons déjà mentionné, nous pourrons étendre de façon cohérente une valuation sur $I^{[2]}$, à une valuation sur $I \times I$, en posant la valeur d'une constante sur la diagonale de $I \times I$.

Conformément au schéma de la figure 1 du paragraphe 1, l'indice d'association brut entre X et Y se met sous la forme

$$< X,Y > = \sum \left\{x_{ij}\, y_{ij} / (i,j) \in K\right\}. \qquad (15)$$

Pour l'hypothèse d'absence de liaison H_1, nous considérons $K = I^{[2]}$; de sorte que l'indice brut aléatoire S_1 prend la forme suivante :

$$s(\sigma,\tau) = < X_1^*, Y_1^* > = \sum \left\{ x_{\sigma(i)\,\sigma(j)}\, y_{\tau(i)\,\tau(j)} / \, (i,j) \in I^{[2]} \right\} \quad (16)$$

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

La distribution de $<X_1^*, Y_1^*>$ est identique à celle commune des deux variables aléatoires duales $<X_1^*, Y>$ et $<X, Y_1^*>$ (voir par exemple [Lerman 1981_b, chap.2].

Pour H_2, nous considérons $K = I \times I$; de sorte que l'indice brut aléatoire S_2 s'exprime comme suit :

$$s(I^*, I'^*) = < X_2^*, Y_2^* > = \sum \left\{ x_{i_j^* i_k^*}\, y_{i'^*_j i'^*_k} / \, 1 \leq j \, , k \leq n \right\} \quad (17)$$

où $$I^* = i_1^*, i_2^*, \ldots, i_j^*, \ldots, i_n^*$$

et $$I'^* = i'^*_1, i'^*_2, \ldots, i'^*_j, \ldots, i'^*_n$$

sont deux suites aléatoires indépendantes, conformément au modèle N_2 [cf.(*ii*),§ 2.1].

Pour H_3, nous considérons $K = I \times I$; de sorte que l'indice brut aléatoire S_3 s'écrit :

$$s\,(I_\nu^*, I'^*_\nu) = < X^*{}_3, Y^*{}_3 > = \sum \left\{ x_{i_j^* i^*_k}\, y_{i'^*_j i'^*_k} / \, 1 \leq j \, , k \leq \nu \right\} \quad (18)$$

où les deux suites aléatoires :

$$I_\nu^* = \left(i_1^*, i_2^*, \ldots, i_j^*, \ldots, i_\nu^* \right)$$

et

$$I'^*_\nu = \left(i'^*_1, i'^*_2, \ldots, i'^*_j, \ldots, i'^*_\nu \right)$$

sont conditionnés par le même entier aléatoire . Pour $\nu = l$, I_l^* et I'^*_l sont deux suites aléatoires indépendantes, conformément au modèle N_3 [cf.(*ii*), § 2.1].

Revenons à l'indice aléatoire S_1 qui seul, nous concernera ici dans la suite. Il est à notre connaissance apparu pour la première fois dans une étude très spécifique de H.E. Daniels [Daniels 1944] concernant l'association entre les coefficients ρ_S de Spearman et τ de M.G. Kendall. G. Lecalvé [Lecalvé 1976] s'en inspira pour proposer une extension de nos coefficients d'association entre variables qualitatives. Nous avons repris cette étude de façon plus précisément combinatoire [Lerman 1977, 1981], ce qui nous a conduit à une expression formelle claire des moments de la distribution de S_1. Nous ignorions alors la contribution de N. Mantel [Mantel 1967] qui - dans une optique de régression - considère la même statistique S_1 et en donne les deux premiers moments $E(S_1)$ et var(S_1). Cependant, la nature formelle de notre expression de var(S_1) est assez différente de celle de N. Mantel. Nous ignorions également l'intérêt de L. Hubert - qui était au courant du travail de N. Mantel - pour la statistique S_1. L. Hubert considère cette dernière dans une optique de test d'hypothèse pour l'épreuve de la conformité entre deux matrices de proximité [Hubert 1983, 1987].

Comme nous l'avons déjà exprimé dans l'introduction générale, notre optique concerne la mise en évidence d'un coefficient d'association entre deux variables relationnelles valuées, à partir de

$$Q(X,Y) = \frac{s - E(S_1)}{\sqrt{var(S_1)}}, \qquad (19)$$

où $s = \langle X,Y \rangle$ [*cf.*(15)] et où E et var désignent respectivement l'espérance mathématique et la variance [*cf.*(5) §1].

Nous avons pu nous rendre compte que dans le cas de la comparaison de deux relations unaires, le coefficient $Q(X,Y)$ peut s'écrire sous la forme :

$$Q(X,Y) = \sqrt{n}\, r(X,Y), \qquad (20)$$

où $r(X,Y)$ est un coefficient indépendant d'un effet de taille, qui a le sens d'un indice de corrélation. Sous des conditions très générales $r(X,Y)$ tend vers sa valeur théorique$\rho(X,Y)$ calculée au niveau d'une population infinie P, lorsque l'ensemble O des objets est un échantillon aléatoire de taille croissante.

La première question fondamentale est de savoir s'il en est de même pour la comparaison de deux variables relationnelles binaires. Dans ce dernier cas, il y a lieu de dégager l'expression formelle de $r(X,Y)$ et étudier son comportement. A cette fin, nous mettrons en évidence une décomposition de $var(S_1)$, en éléments positifs, chacun d'un ordre fixé par rapport à n. Il s'agira ensuite de "voir" cette expression limite dans différents cas de figure, relevant notamment de la comparaison de variables qualitatives.

$r(X,Y)$ a ainsi le sens d'un coefficient de corrélation ; c'est-à-dire, en d'autres termes, de mesure de la concomitance des deux relations X et Y. On peut d'autre part remarquer que dans le cas unaire et quelle que soit l'hypothèse d'absence de liaison H_i (i=1,2 ou 3), le coefficient

$$r'_i(X,Y) = \frac{Q_i(X,Y)}{\sqrt{Q_i(X,X) \times Q_i(Y,Y)}} \qquad (21)$$

n'est autre que celui de K. Pearson dans le cas discret (booléen) et celui de Bravais-Pearson dans le cas valué (numérique). Dans ces conditions et en tenant compte de ce que la réponse à la question posée est positive, on peut considérer les deux types de coefficients $r(X,Y)$ et $r'(X,Y)$ dans le cas binaire et même dans le cas q-aire, où q est un entier fixé quelconque.

Comme nous l'avons mentionné dans l'introduction générale, nous allons étudier le cas où les deux valuations X et Y sont toutes les deux, soit symétriques, soit antisymétriques [cf. (6) et (7), § 1]. Nous considérerons ensuite l'application des résultats obtenus à deux cas qualitatifs importants et de base ; celui nominal et celui, totalement ordinal. Pour ce dernier, on envisagera également une représentation qui ne correspond pas à un codage antisymétrique tel que celui (7) du paragraphe 1. On cherchera d'autre part, à présenter une extension relativement à une situation originale de comparaison de deux relations q-aires. Pour conclure, nous ferons état de travaux très récents menés dans le cadre de thèses qui poursuivent et développent le présent travail.

Considérons ici l'indice brut centré [s-$E(S_1)$], numérateur de $Q(X,Y)$ [cf(19)]. On a

$$E(S_1) = n^{[2]}\, \mu_X\, \mu_Y, \qquad (22)$$

où $n^{[2]} = n(n\text{-}1)$ et où μ_X (resp. μ_Y) représente la moyenne de la valuation X (resp. Y) sur $I^{[2]}$. Plus précisément, on a

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

$$\left(resp. \mu_Y = \frac{1}{n^{[2]}} \sum \left\{ y_{ij} / (i,j) \in I^{[2]} \right\} \right) \qquad (23)$$

BIBLIOGRAPHIE

[1] ARABIE P. and HUBERT L.J., (1992), "Combinatorial Data Analysis" 1992, *Annual Review of Psychology* (to appear).

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

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

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

[5] DAUDE F., (1990), "Normalisation sous hypothèse d'absence de lien", *Publication interne* Irisa, Rennes, n°°549, septembre 1990, 42 pages.

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

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

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

[9] GOODMAN L.A. and KRUSKAL W.H., (1963), "Measures of association for cross classifications III : Approximate sampling theory", Vol. 58, pp. 310-364.

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

[11] HUBERT L.J., (1983), "Inference procedures for the evaluation and comparison of proximity matrices", in Numerical Taxonomy, Ed. J. Felsenstein, NATO ASI Series, Springer Verlag.

[12] HUBERT L.J., (1987), "Assignment methods in combinatorial data analysis", Marcel Decker, New York, 1987.

[13] KENDALL M.G., (1970), "Rank correlation methods", Charles Griffin, fourth edition (first edition in 1948).

[14] LECALVÉ G., (1976), "Un indice de similarité pour des variables de types quelconques", *Statistique et Analyse des Données*, 01-02, pp. 39-47.

[15] LERMAN I.C. (1973), "Etude distributionnelle de statistiques de proximité entre structures finies de même type ; application à la classification automatique", *Cahiers du Buro*, n° 19, Paris.

[16] LERMAN I.C., (1976), "Formal analysis of a general notion of proximity between variables", Congrès Européen des Statisticiens, Grenoble, Sept. 1976, North Holland (1977).

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

[18] LERMAN I.C., (1983), "Association entre variables qualitatives ordinales nettes ou floues", *Statistique et Analyse des données*, vol. 8 n°7, pp. 41-73.

[19] LERMAN I.C., (1984), "Justification et validité statistique d'une échelle [0,1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées", *Publ. Inst. Stat. Univ. Paris XXIX*, fasc. 3-4, pp. 27-57.

[20] LERMAN I.C., (1987a), "Construction d'un indice de similarité entre objets décrits par des variables d'un type quelconque. Application au problème du consensus en classification", *Rev. Statistique Appliquée*, XXXV (2), pp. 39-60.

[21] LERMAN I.C. (1987_b), "Analyse de la forme limite de coefficients statistiques d'association entre variables relationnelles", *Rapport de recherche* n° 702, Inria, Juillet 1987.

[22] LERMAN I.C., (1987_c), "Maximisation de l'association entre deux variables qualitatives ordinales", *Rev. math. Sci. hum.*, 25ème année, n° 100, 1987, pp. 49-56.

[23] LERMAN I.C., (1988), "Structure maximale pour la somme des carrés d'une contingence aux marges fixées; une solution algorithmique programmée", *Rairo*, vol. 22, n°2, pp. 83 à 136.

[24] LERMAN I.C., (1991), "Foundations of the Likelihood Linkage Analysis (LLA) Classification method", *Applied Stochastic Models and Data Analysis*, vol. 7, pp. 63-76 (J. Wiley).

[25] LERMAN I.C. et GHAZZALI N., (1991), "Quoi retenir d'un arbre de classification ? Un essai en quantification d'image numérisée", *Rapport de recherche n° 1386*, Inria, Janvier 1991.

[26] LERMAN I.C., GRAS R. et ROSTAM H., (1981), "Elaboration et évaluation d'un indice d'implication pour des données binaires" I et II ; I : *Math. Sci. hum.*, 19ème année, n° 74, 1981 pp. 5-35, II : *Math. Sc. hum.*, 19ème année, n° 75, 1981, pp. 5-47.

[27] LERMAN I.C. et PETER Ph., (1985), "Organisation et consultation d'une banque de petites annonces" à partir d'une méthode de classification hiérarchique en parallèle", *Journées Internationales Analyse des Données et Informatique IV*, Octobre 1985, Versailles, North Holland (1986), pp. 121-136.

[28] LERMAN I.C. et PETER Ph., (1989), "Classification of concepts described by taxonomic preordonnance variables with multiple choice. Application to the structuration of a species set of phebotomine" *in* Data Analysis, Learning symbolic and numerical knowledge, edited by E. Diday, Inria, Nova Science Publishers, (1989), pp. 73-87.

[29] MANTEL N., (1967), "Detection of disease clustering and a generalized regression approach", *Cancer Research*, vol. 27, n° 2, pp. 209-220.

[30] MESSATFA H., (1990), "Unification relationnelle des critères et structures optimales des tables de contingences", thèse de doctorat de l'Université Pierre et Marie Curie, 5 mars 1990.

[31] MIELKE W., (1979), "On asymptotic non normality of null distributions of MRPP Statistics", *Communications in Statistics, Theory and Methods*, A8 (15), pp. 1541-1550.

[32] NOETHER G., (1949), "On a theorem by Wald and Wolfowitz", *Ann. Math. Stat.* vol. 20, pp. 455-458.

[33] OUALI-ALLAH M., (1991_a), "Analyse en préordonnances des données qualitatives. Applications aux données numériques et symboliques", Thèse de l'Université de Rennes I, 5 décembre 1991, Rennes, Université de Rennes I.

[34] OUALI-ALLAH M, (1991_b), "Avare : un programme de calcul des associations entre variables relationnelles", *Publication interne Irisa*, n° 591, juin 1991, 32 pages.

[35] PETER Ph. (1987), "Méthodes de classification hiérarchique et problèmes de structuration et de recherche d'informations, assistées par ordinateur", thèse de l'Université de Rennes I, 6 mars 1987, Rennes, Université de Rennes I.

[36] SUPPES P. and ZINNES J.L., (1963), "Basic measurement theory" in Handbook of mathematical psychology, Eds Bush. Luce, Galanter, New York, J. Wiley, pp. 2-76.

[37] TARSKI A., (1954), "Contribution to the theory of models", I, II. *Indagationes Mathematicae*, 16, pp. 572-588.

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