JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

I. C. LERMAN

Rôle de l'inférence statistique dans une approche de l'analyse classificatoire des données

Journal de la société statistique de Paris, tome 127, nº 4 (1986), p. 238-252

<http://www.numdam.org/item?id=JSFS_1986__127_4_238_0>

© Société de statistique de Paris, 1986, tous droits réservés.

L'accès aux archives de la revue « Journal de la société statistique de Paris » (http://publications-sfds.math.cnrs.fr/index.php/J-SFdS) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

RÔLE DE L'INFÉRENCE STATISTIQUE DANS UNE APPROCHE DE L'ANALYSE CLASSIFICATOIRE DES DONNÉES

I.C. LERMAN
I.R.I.S.A. (1)

Nous montrons comment une approche de la classification hiérarchique — basée sur la « vraisemblance des liaisons observées » — intègre et développe de façon spécifique, conformément à l'optique de l'analyse des données, certains aspects d'une statistique non paramétrique et combinatoire. Cette optique est située par rapport à celle des tests d'indépendance ainsi que par rapport au point de vue inférentiel où l'ensemble des objets sur lequel se base l'étude est regardé comme un échantillon aléatoire d'une vaste population.

We show how an approach of hierarchical classification — based on the "likelihood of observed relations" — involves and develops in a specific way different aspects of combinatorial and non parametric statistics. This developement is done according to Data Analysis principles. These last principles are situated with respect to those of independence hypotheses tests. On the other hand, we analyze the approach from inferential point of view where the set of objects is considered as a random sample from a large population.

I — INTRODUCTION

Un reproche constant des Statisticiens « classiques » vis-à-vis des Analystes des Données porte sur la non considération par ces derniers de la population mère $\mathcal{T}$ dont provient l'échantillon « aléatoire » E sur lequel se base l'étude et les conclusions. Nous désignerons par N la taille — supposée « très grande » — de $\mathcal{T}$ et par n la taille de E.

Notre objectif ici consiste à analyser notre approche de la classification d'une famille de variables observées sur E, par rapport à ce point de vue inférentiel où E est la réalisation d'un échantillon aléatoire de $\mathcal{T}$. Pour mener à bien notre réflexion, nous considérerons le cas le plus simple où les variables sont soit des attributs logiques de description (variables de présence-absence), soit des variables quantitatives.

Relativement à ce problème des liaisons mutuelles entre variables descriptives de $\mathcal{T}$ observées sur E, on peut se référer à deux optiques :

(i) « Tests de l'hypothèse d'indépendances mutuelles ».
(ii) « Analyse des Données ».

Ces deux approches sont en fait dans leurs philosophies respectives mentalement opposées. En effet, la première (celle des « tests d'hypothèses ») privilégie la croyance en l'absence de liaisons, lesquelles, lorsqu'elles se trouvent quand même établies — sur la base de E et avec un seuil fixé — ne peuvent être réellement « mesurées » et comparées.

Au contraire, pour l'« analyse des données », il n'y a aucun doute quant à l'existence des liaisons entre les variables sur $\mathcal{T}$, cependant ces liens sont plus ou moins forts ou plus ou moins ténus et il s'agit de les évaluer de façon objective pour les organiser au mieux. C'est de par la clarté de la compréhension de l'interprétation des résultats que l'induction au niveau de la population parente se fait de façon naturellement implicite et ce, sans se poser des questions sur la qualité des estimations calculées — sur la base de l'échantillon E — des indices d'associations entre variables.

Pour mettre en évidence la contradiction logique entre les deux approches, considérons l'exemple très simple suivant : a, b, c et d sont quatre attributs mutuellement distincts définis pour la

(1) I.R.I.S.A. Institut de Recherche en Informatique et Systèmes aléatoires, Université de Rennes, Campus universitaire de Beaulieu, Avenue du Général-Leclerc, 35042 RENNES Cedex

Journal de la Société de statistique de Paris, tome 127, nº 4, 1986.

description d'une population $\mathfrak{T}$ $\pi(a)$, $\pi(b)$ et $\pi(a \wedge b)$ (*) [resp. $\pi(c)$, $\pi(d)$ et $\pi(c \wedge d)$] désignent respectivement — au niveau de $\mathfrak{T}$ — les proportions de sujets où a, b et $a \wedge b$ (resp. c, d et $c \wedge d$) se trouvent présents.

On suppose que le couple (a,b) a été observé sur un échantillon E — extrait selon les lois du sondage aléatoire — de taille $n = 100$ et que la situation est la suivante :

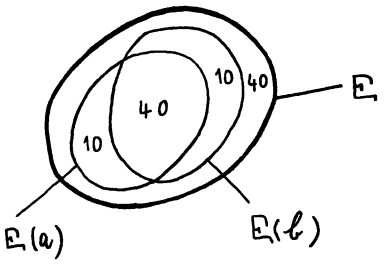

Figure 1

On suppose d'autre part, que le couple (c,d) a été observé sur un échantillon F de taille $m = 10\,000$ et que la situation est la suivante :

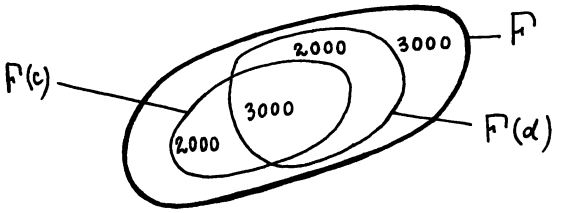

Figure 2

Désignons par $p(a)$, $p(b)$ et $p(a \wedge b)$ [resp. $p(c)$, $p(d)$ et $p(c \wedge d)$] les proportions de présence des attributs a,b et $a \wedge b$ (resp. c,d et $c \wedge d$) au niveau de E et de F, respectivement. D'autre part, désignons par $P(a)$, $P(b)$ et $P(a \wedge b)$ [resp. $P(c)$, $P(d)$ et $P(c \wedge d)$] les proportions aléatoires de présence des attributs a, b et $a \wedge b$ (resp. c, d et $c \wedge d$) au niveau d'un échantillon aléatoire $\mathcal{E}$ de taille 100 (resp. $\mathcal{F}$ de taille 10^4). Dans l'hypothèse d'indépendance $\pi(a \wedge b) = \pi(a)\pi(b)$ [resp. $\pi(c \wedge d) = \pi(c)\pi(d)$]

$\sqrt{n}\, R(a, b) = \sqrt{n}\, [P(a \wedge b) - P(a)\, P(b)] / \sqrt{P(a)\, P(b)}$
{resp. $\sqrt{m}\, R(c, d) = \sqrt{m}\, [P(c \wedge d) - P(c)\, P(d)] / \sqrt{P(c)\, P(d)}$} suit une loi normale centrée réduite $N(0,1)$.

Considérons dans ces conditions le test de l'hypothèse d'indépendance entre a et b d'une part, c et d d'autre part, au même seuil $\alpha = 0{,}001$.

Le calcul donne

$$r(a, b) = [p(a \wedge b) - p(a)p(b)] / \sqrt{p(a)p(b)} = 0{,}3$$

et

$$r(c, d) = [p(c \wedge d) - p(c)p(d)] / \sqrt{p(c)p(d)} = 0{,}1$$

D'où

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

(*) a ∧ b indique la conjonction de a et de b.

Par conséquent — au seuil $\alpha = 0{,}001$ choisi — l'hypothèse d'indépendance est violemment rejetée pour ce qui concerne la comparaison entre c et d, mais ne peut être rejetée pour ce qui concerne la comparaison entre a et b.

A partir de là, le chemin est court dans l'esprit de l'utilisateur pour arriver à croire que « a et b ne sont peut-être pas très liés, mais c et d le sont très certainement ».

Pour contredire cette croyance, l'analyse des données fera le pari d'une valeur 0,1 du cœfficient théorique.

$$\rho(c, d) = [\pi(c \wedge d) - \pi(c)\,\pi(d)] / \sqrt{\pi(c)\,\pi(d)}.$$

Dans le cadre d'un tel pari — qui ne paraît pas trop risqué puisque $r(c,d) = 0{,}1$ sur la base d'un échantillon de taille 10^4 — ont a (cf. [14] où ce résultat est établi de façon précise)

$$Pr\{R(a, b) \geqslant 0.3/\rho(a, b) \leqslant \rho(c, d) = 0.1\} \leqslant 10^{-5}\,!$$

Cependant, la critique du statisticien classique garde toute sa pertinence. On a en effet un problème de stabilité extrinsèque puisque E n'est que la réalisation d'un échantillon aléatoire de $\mathfrak{T}$. Ce problème de stabilité est crucial pour un effectif n de l'échantillon non « assez grand » ou pour les liaisons « faibles ». Dans ce cas en effet, les fluctuations d'échantillonnage peuvent entraîner l'apparition d'aberrations. Nous allons à cet égard, rappeler une expérience statistique qui a été effectuée dans le cadre d'un D.E.A. (cf. [2]).

Il s'agit de l'étude expérimentale de la stabilité de notre classification hiérarchique appliquée à un ensemble d'attributs (*i.e.* variables logiques de présence-absence) lorsque l'échantillon des individus, défini selon un mode aléatoire, croît. C'était à divers titres une situation idéale pour une telle étude : la famille d'attributs résulte d'un questionnaire qu'on remplit lors de l'établissement d'un « bilan de santé » par un centre d'examen de santé de la Sécurité Sociale (il s'agit en l'occurrence de celui de Rennes), certaines liaisons entre les attributs pouvaient être fortes mais la plupart étaient ténues. Toutefois, disposant de près de 14 000 bilans par an, on pouvait à sa guise faire croître l'échantillon observé de la population consultante du centre. Enfin, deux traitements parallèles ont été analysés et concernant respectivement les populations masculine et féminine.

Pour chacune des deux populations, nous avons considéré une suite croissante d'échantillons aléatoires dont la suite des tailles est la suite des multiples entiers de 1 000 (1 000, 2 000, 3 000,...). On a pu constater que les classes hiérarchiques d'attributs qui étaient bien structurées [marquées par des « nœuds significatifs » (cf. ci-dessous)] et correspondaient à des liaisons nettes, se trouvaient préservées. D'autres classes correspondantes à des profil plus flous pouvaient au départ comprendre des éléments aberrants, mais l'échantillon des individus augmentant, ces éléments aberrants se déplaçaient pour donner meilleure consistance à d'autres classes, alors que les classes abandonnées devenaient plus cohérentes en s'adjoignant parfois et de façon compatible des attributs restés isolés lors d'un précédent traitement (pour un échantillon de taille plus petite). Au bout de $n = 4000$ pour la population des « hommes » (resp. de $n = 3000$ pour la population des « femmes ») la stabilité parfaite se trouvait atteinte; en d'autres termes, la classification hiérarchique des attributs restait invariable lorsqu'on augmentait la taille de l'échantillon. Le fait que la stabilité ait été atteinte plus rapidement pour la population féminine que pour celle masculine semble dénoter une moins grande dispersion comportementale pour les femmes. Nous avons pu nous rendre compte que la rapidité de la convergence vers la stabilité dépendait de deux facteurs : le premier peut être défini par la force et la séparabilité des tendances comportementales sous-jacentes (« classificabilité » de l'ensemble des variables-attributs [Lerman (1970*a*), (1981)] et le second facteur — on pouvait s'en douter — est lié à la fréquence de présence des attributs; en effet, la stabilité se trouve affectée par les attributs rares.

De sorte qu'il est difficile de répondre à la question de l'utilisateur qui demande quelle est la taille de l'échantillon qu'il lui faut pour extraire les tendances comportementales de la population qu'il étudie. En effet, tout dépend de la force et de la séparabilité de ces tendances, d'autant plus que la

vérité générale est que certains profils de comportement sont bien marqués et d'autres le sont moins. Toutefois, dans le calcul des indices d'association entre variables que nous lui proposerons, nous limiterons le nombre de chiffres significatifs, non pas en fonction de la précision de calcul de l'ordinateur, mais en fonction de la taille de l'échantillon. Une telle limitation peut conduire — lorsque la taille de l'échantillon n'est pas suffisante — à isoler dans la structure hiérarchique des classifications sur l'ensemble des variables, des éléments qui se seraient accrochés de façon artificielle aux classes réelles.

II — Distributions non paramétriques et analyse des données

La conception des méthodes de l'analyse des données se situe — nous l'avons dit — au niveau de l'échantillon E sans que des hypothèses formulées au niveau de la population $\mathcal{P}$ aient à intervenir directement. Il ne faut pas croire qu'il s'agit là d'une caractéristique de distinction par rapport aux méthodes de la statistique inductive. En effet, un aspect important des tests non-paramétriques d'hypothèses se conçoit uniquement au niveau de E; ainsi en est-il des tests de permutation où nous nous contenterons de citer une des publications pionnières [Wald & Wolfowtiz (1944)]. Dans ce dernier test, pour établir le lien entre deux variables numériques v et w, on associe à la suite des valeurs observées sur E $(x_1, x_2, ..., x_i, ...x_n)$ de la variable v [resp. $(y_1, y_2, ..., y_i, ..., y_n)$ de la variable w] — où $\{1, 2, ..., i, ..., n\}$ code E — la suite aléatoire $(x_{\sigma(1)}, x_{\sigma(2)}, ..., x_{\sigma(i)}, ..., x_{\sigma(n)})$, [resp. $(y_{\tau(1)}, y_{\tau(2)}, ..., y_{\tau(i)}, ..., y_{\tau(n)})$] où σ (resp. τ) est un élément aléatoire dans l'ensemble Gn, muni d'une probabilité uniforme des n! permutations sur $(1, 2, ..., i, ..., n)$. On situe ensuite l'indice $\sum\{x_i y_i / 1 \leqslant i \leqslant n\}$ (que nous appelons « brut ») par rapport à la distribution commune de l'une ou de l'autre des deux variables aléatoires (*v.a.*) :

$\sum\{x_i y_{\tau(i)} / 1 \leqslant i \leqslant n\}$ et $\sum\{x_{\sigma(i)} y_i / 1 \leqslant i \leqslant n\}$. L'association $(1, 2, ..., i, ..., n) \rightarrow (\sigma(1), \sigma(2), ..., \sigma(i))$ (resp. $(\tau(1), \tau(2), ..., \tau(i), ..., \tau(n))$ est ce que nous désignons par « hypothèse d'absence de liaison » *(h.a.l.)*.

Posons $s(v, w) = \sum\{x_i y_i / 1 \leqslant i \leqslant n\}$, $s(v, w^*) = \sum\{x_i y_{\tau(i)} / 1 \leqslant i \leqslant n\}$, $s(v^*, w) = \sum\{x_{\sigma(i)}\, y_i / 1 \leqslant i \leqslant n\}$ et $s(v^*, w^*) = \sum\{x_{\sigma(i)} y_{\tau(i)} / 1 \leqslant i \leqslant n\}$, où σ et τ sont deux *permutations aléatoires indépendantes. Il est aisé de calculer la moyenne et la variance* μ_{vw} et σ^2_{vw} de la loi commune de $s(v, w^*)$, $s(v^*, w)$ et $s(v^*, w^*)$, qui est — sous des conditions assez générales — asymptomatiquement normale.

Si maintenant nous considérons l'indice « brut $s(v, w)$ centré et réduit par rapport à l'hypothèse d'absence de liaison, on a

$$[s(v, w) - \mu_{vw}] / \sqrt{\sigma^2_{vw}} = \sqrt{(n-1)}\, r_{vw}, \quad (1)$$

où r_{vw} n'est autre que le coefficient de corrélation de Bravais Pearson.

Ainsi, cette méthode de « standardization » permet *la découverte de l'expression formelle de l'indice de corrélation dont l'introduction n'est nullement liée au caractère plus ou moins linéaire de la relation entre les deux variables.*

Maintenant, ayant rejeté l'optique des tests de l'hypothèse d'indépendance, on peut récupérer l'échelle de probabilité définie à partir de la loi normale d'approximation, pour « mesurer » le degré de la liaison entre les deux variables à partir de l'indice suivant de la « vraisemblance du lien » (« v. l. ») :

$$J(v, w) = Pr\{s(v^*, w^*) \leqslant s(v, w)\} \cong \Phi[\sqrt{(n-1)}\, r_{vw}], \quad (2)$$

où les deux variables v et w sont considérées d'autant plus liées que $s(v,w)$ est invraisemblablement grand, par rapport à l'hypothèse d'absence de liaison. Φ désigne la f.r. de loi N(0,1).

Cependant, on remarque qu'en cas de non parfaite indépendance ($\rho_{vw} \neq 0$ au niveau de la population mère), l'indice « v. l. » tel qu'il a été défini avoisine la valeur $+1$ (si $\rho_{vw} > 0$) ou celle 0 (si $\rho_{vw} < 0$, pour n suffisamment grand. Est-ce le retour du point de vue inférenfiel qui nous interdit d'utiliser l'échelle de probabilité?

En réalité, en analyse des données, le problème n'est pas tant de « mesurer » l'association entre deux variables seulement, mais — comme nous l'avons dit ci-dessus — d'organiser de façon mutuelle les liaisons concernant une famille nombreuse de variables. D'ailleurs, on peut même exprimer que si notre univers se limitait aux deux seules variables v et w, on peut proposer n'importe quel nombre pour la mesure de l'association entre les deux variables! Dans ces conditions, pourquoi pas la valeur donnée par l'expression (2). Toutefois, on attendra d'aborder la comparaison deux à deux d'une famille de variables pour voir comment utiliser opérationnellement l'échelle de probabilité obtenue à partir de la vraisemblance des liaisons observées.

De la même manière que nous avons considéré la comparaison de deux variables quantitatives, nous allons à présent envisager la comparaison de deux attributs logiques. Dans ce cas, le langage permutationnel sera avec économie remplacé par un langage ensembliste.

Si (a,b) est un couple d'attributs descriptifs, on a la représentation suivante de la situation relative entre les deux attributs :

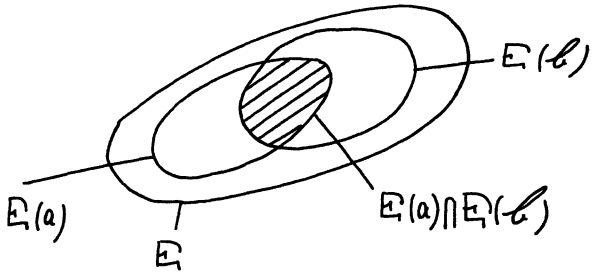

Figure 3

L'indice « brut » d'association est

$$s(a, b) = \text{card}\,[E(a) \cap E(b)]. \quad (3)$$

Les deux indices aléatoires duaux et de même loi, se mettent respectivement sous la forme

$$s(a, b^*) = \text{card}\,[E(a) \cap Y] \text{ et } s(a^*, b) = \text{card}\,[X \cap E(b)], \quad (4)$$

où X (resp. Y) est un élément aléatoire pris uniformément au hasard dans l'ensemble $\mathfrak{P}_{n(a)}(E)$ [resp. $\mathfrak{P}_{n(b)}(E)$] des parties de E de même cardinal $n(a)$ [resp. n(b)].

On pourra introduire de façon plus symétrique l'indice aléatoire $s(a^*, b^*) = \text{card}\,(X \cap Y)$, de même loi que $s\,(a^*, b)$ [resp. $s\,(a^*, b)$], où X et Y sont choisis comme il est exprimé ci-dessus et indépendamment l'un de l'autre. μ_{ab} et σ^2_{ab} désignant la moyenne et la variance de l'indice brut aléatoire, l'indice brut réduit prend la forme suivante :

$$Q(a, b) = [s(a, b) - \mu_{ab}]/\sqrt{\sigma^2_{ab}} = \sqrt{(n-1)}\,r_{ab}, \quad (5)$$

où r_{ab} n'est autre que le coefficient d'association de K. Pearson.

Comme dans le cas de la comparaison de deux variables quantitatives, on voit comment — par « standardization » — *l'hypothèse d'absence de liaison permet la découverte de l'expression formelle d'un indice de type corrélatif.* Ici encore, on peut introduire, mais sans encore un effet opérationnel, l'indice de la vraisemblance du lien :

$$\mathfrak{I}(a, b) = Pr\{s(a^*, b^*) \leqslant s(a, b)\} \cong \Phi\,[\sqrt{(n-1)}\,r(a, b)]. \quad (6)$$

Nous avons généralisé extensivement cette approche pour la comparaison de deux variables relationnelles de n'importe quel type (cf. [Lerman (1973) (1981), Chap. 2]). Mais pour être fidèle, la représentation ensembliste de telles variables ne peut plus se faire au niveau de l'ensemble E des objets. Elle doit se faire au niveau de $E \times E$, voire même au niveau de $(E \times E) \times (E \times E)$. Le schéma général est le suivant :

$$\begin{aligned}
(\alpha, \beta) &\longrightarrow [R(\alpha), R(\beta)] \in \Omega_1 \, X \, \Omega_2 \\
S(\alpha, \beta) &= \operatorname{card}[R(\alpha) \cap R(\beta)] \\
&\text{h.a.l.} \\
S &= s(\alpha^*, \beta^*) = \operatorname{card}[R(\alpha^*) \cap R(\beta^*)] \\
Q(\alpha, \beta) &= [s(\alpha, \beta) - \mathcal{E}(S)] / \sqrt{\operatorname{var.}(S)} \\
&= \sqrt{n}\, r(\alpha, \beta) \\
\mathfrak{I}(\alpha, \beta) &= \Phi[Q(\alpha, \beta)]. \qquad (7)
\end{aligned}$$

Pour illustrer ce schéma, considérons le cas de la comparaison de deux variables qualitatives nominales (c'est un des cas les plus simples de variables relationnelles). Supposons que α (resp. β) définisse une partition de type $t = (n_1, n_2, ..., n_k)$ [resp. $s = (m_1, m_2, ..., m_h)$] : n_i $(1 \leqslant i \leqslant k)$ [resp. m_j $(1 \leqslant j \leqslant h)$] est le cardinal de la i-ème (resp.j-ème) classe définie par α (resp. β).

$R(\alpha)$ [resp. $R(\beta)$] est l'ensemble des paires d'objets de E réunies par la partition définie par α (resp. β) que également nous notons α (resp. β). Ω_1 (resp. Ω_2) est l'ensemble des parties de $P_2(E)$ — lequel étant l'ensemble des paires d'objets de E — qui représentent une partition de type t (resp. s). α^* (resp. β^*) est une partition aléatoire dans l'ensemble — muni d'une probabilité uniforme — $\mathfrak{T}(n;t)$ [resp. $\mathfrak{T}(n;s)$] des partitions sur E de type t (resp. s). α^* et β^* sont indépendantes.

Revenons un instant sur l'indice $\mathfrak{I}(a, b)$ de la « vraisemblance du lien » dans le cas de la comparaison de deux attributs de description a et b. Ce dernier se met sous la forme

$$\mathfrak{I}(a, b) = Pr\{s(a^*, b^*) \leqslant s(a, b) / \text{h.a.l.}\}. \quad (8)$$

Cet indice est — si on emprunte le vocabulaire de M. Allais [Allais(1983)] — une « fréquence mathématique ». Il s'agit en effet, dans le cadre de l'h.a.l. ci-dessus présentée, de la proportion de couples de parties (X, Y) — avec card $(X) = n(a)$, card $(Y) = n(b)$ — dont le cardinal de l'intersection est inférieur à $s(a,b) = \operatorname{card}[E(a) \cap E(b)]$. Par ailleurs, il faut savoir qu'il y a d'autres formes de l'h.a.l. qui correspondent à munir l'ensemble des parties d'un ensemble d'une mesure de probabilité adéquate pour le choix de X (resp. Y).

Résumons-nous. Nous avons pu nous rendre compte que l'h.a.l. permet la découverte de l'expression formelle d'un indice d'association entre variables relationnelles. C'est ainsi que nous retrouvons les coefficients de K. Pearson et de M.G. Kendall (cf. [7]) et que nous découvrons une classe très riche d'indices d'association. D'autre part, l'indice de la vraisemblance du lien pour la comparaison de deux variables seulement peut certes être proposé, mais ne peut avoir à ce niveau de caractère opérationnel. On peut certes proposer la table de cet indice directement pour les comparaisons deux à deux d'un ensemble V de variables descriptives. Mais, la taille n de l'échantillon augmentant, les valeurs de cette table — dans l'optique de l'analyse des données — vont tendre soit vers zéro, soit vers un. C'est que, pour établir la valeur de l'indice d'association de deux variables, l'h.a.l. ne tient absolument pas compte du contexte des autres variables.

Pour en tenir compte et en nous plaçant pour simplifier dans la cas où la donnée est une famille $\mathcal{A}$ d'attributs logiques de description, nous proposons de remplacer la table suivante des indices — localement centrés et réduits — d'association

$$\{Q(a, b) / \{a, b\} \in P_2(\mathcal{A})\}, \quad (9)$$

[$P_2(\mathcal{A})$ est l'ensemble des paires d'éléments de $\mathcal{A}$], par celle des indices qui sont en plus globalement réduits :

$$\{Q_s(a, b)/\{a, b\} \in P_2(\mathcal{A})\}, \quad (9s)$$

où

$$Q_s(a, b) = Q(a, b)/\sqrt{M_2(Q)},$$

où $M_2(Q)$ est le moment absolu d'ordre 2 de (9).

Q_s (a,b) est d'un point de vue formel une « densité orientée » en (a,b), puisque cet indice rapporte $Q(a,b)$ à $\sqrt{M_2(Q)}$.

Une autre forme de réduction globale basée sur les quantiles de la distribution normale est fournie dans [M.H. Nicolaü (1972)]. Plus précisément, on détermine un paramètre λ positif tel que

$$\max\{|Q(a, b)|/\lambda \ /\{a, b\} \in P_2(\mathcal{A})\} \leqslant 2.5$$

et on remplace la table (9) par celle

$$\{Q_\lambda(a, b) = [Q(a, b)]/\lambda|\{a, b\} \in P_2(\mathcal{A})\}. \quad (9_\lambda)$$

Toutefois, c'est la réduction par le moment d'ordre 2 qui nous a semblé — dans la pratique de la plupart des cas — fournir les résultats les plus cohérents dans leurs nuances pour l'organisation des liens « faibles », lorsqu'on se rapproche des derniers niveaux de l'arbre des classifications. C'est à cette dernière réduction que nous nous référons ci-dessous.

Maintenant, la justification statistique de la référence à une échelle de probabilité doit s'effectuer de façon globale où, à la famille $\mathcal{A}$ des attributs observés, on associe une famille $\mathcal{A}^*$ d'attributs aléatoires et indépendants. Cette association se fait ici de façon non paramétrique et intrinsèque à E. Dans ce contexte, nous nous rendons compte [Lerman (1984)] que la table des indices aléatoires

$$\{Q_s(a^*, b)/\{a^*, b^*\} \in P_2(\mathcal{A}^*)\}, \quad (10)$$

suit asymptotiquement une loi multinormale et que Q_s (a^*,b^*) suit asymptotiquement une loi normale centrée réduite. Ainsi, la table des indices que nous retenons définitivement est

$$\{P_s(a, b) = \Phi[Q_s(a, b)]/\{a, b\} \in P_2(\mathcal{A})\}, \quad (11)$$

cette dernière utilise pleinement le pouvoir discriminant de l'échelle de probabilité.

Nous avons dès le départ posé le problème de l'organisation des liaisons mutuelles entre variables. Nous proposons quant à nous pour une telle organisation un arbre hiérarchique condensé à ses nœuds les plus « significatifs ».

A toutes fins utiles et entre parenthèses, nous pouvons signaler que nous disposons d'une approche duale pour aboutir à une table d'indices telle que (11); mais cette fois-ci, pour comparer de façon mutuelle un ensemble d'objets décrits de façon quelconque [Lerman-Peter(1985)].

III — Construction de l'arbre et reconnaissance statistique de ses composants les plus significatifs

Le principe de la construction d'un arbre de classification sur un ensemble fini A peut sembler trivial. Il suffit en effet de se donner une notion de « proximité » ou de « distance » entre parties

disjointes de A. On partira de la partition la plus fine et on réunira à chaque pas les paires de classes qui réalisent la plus grande proximité (ou la plus petite distance).

Il faut cependant savoir que *tout, mais alors tout, est dans la notion de proximité (ou distance),* compte tenu de ce que représente A. Dans notre cas, nous partons de la table (11) ci-dessus pour comparer selon la même démarche deux parties disjointes B et C de $\mathcal{A}$. Ainsi, à la famille des indices

$$\{P_s(b, c)/(b, c) \in B \times C\}, \quad (12)$$

nous associons

$$\{P_s(b^*, c^*)/(b^*, c^*) \in B^* \times C^*\}, \quad (13)$$

où B^* (resp. C^*) est un ensemble d'attributs aléatoires indépendants qu'on associe de façon adéquate à B (resp. C). Cette association se fait — comme ci-dessus — sans référence à une population mère.

La loi du maximum de la famille (13) de *v.a.* conduit au critère de la « vraisemblance du lien maximal » [Lerman (1970b)]. La loi de la moyenne de cette famille conduit au critère de la « vraisemblance de la moyenne » [F. Nicolaü (1980)].

Cependant, il y a une contrainte inhérente à tout algorithme de classification ascendante hiérarchique dans la formation des classes et sous-classes qui sont sous-jacentes au comportement de la population étudiée. En effet, un tel algorithme opère par fusion *d'une paire* de classes à la fois et ce, même si plusieurs paires de classes se trouvent réunies au même niveau de l'arbre parce qu'elles réalisent « en même temps » la plus grande valeur de l'indice d'association entre classes. Or, certaines associations déterminent des compléments de classes correspondantes à un certain niveau de synthèse. D'autre part, certains niveaux correspondent à des états d'équilibre dans la synthèse.

Pour détecter ces niveaux « significatifs » et ces nœuds « significatifs » (correspondants à l'achèvement de classes) nous introduisons un critère qui évalue « combien » les inégalités entre paires d'éléments singletons — conformément à la valeur de l'indice d'association — se trouvent préservées dans la représentation selon l'arbre de classification produit. Il s'agit d'un critère basé sur l'« ordonnance » $\omega(A)$: ordre (ou préordre) total sur l'ensemble $F = P_2(A)$ des paires (ou parties à deux éléments) de A, qu'on peut supposer établi de telle sorte qu'une paire ait un rang d'autant plus élevé que la ressemblance entre ses composantes est plus grande. Ainsi, si Q désigne l'indice de proximité ou d'association sur A, on pose

$$[\forall(p, q) \in F \times F], p<q \iff Q(a, b) < Q(c, d), \quad (14)$$

où on a posé $p = \{a,b\}$ et $q = \{c,d\}$.

On représente $\omega(A)$ au niveau de $F \times F$, de façon ensembliste, au moyen de

$$\mathrm{gr}(\omega) = \{(p, q)/(p, q) \in F \times F, (p<q) \wedge [\neg(q<p)] \text{ pour } \omega\}, \quad (15)$$

où l'expression $(p<q) \wedge [\neg(q<p)]$signifie p précède q et non q précède p.

De la même façon, une partition π sur A (éventuellement délimitée par un même niveau de l'arbre des classifications) est considérée comme définissant un préordre total à deux classes sur F : la première $S(\pi)$ est définie par l'ensemble des paires séparées par la partition et la seconde $R(\pi)$, par l'ensemble des paires réunies par la partition.

Pour ce dernier préordre

$$(\forall(p, q) \in F \times F), (p<q) \wedge [\neg(q<p)] \Leftrightarrow (p, q) \in S(\pi) \times R(\pi). \quad (16)$$

Ainsi, la représentation ensembliste de π au niveau de $F \times F$, se fait au moyen du rectangle :

$$S(\pi) \times R(\pi). \quad (17)$$

Maintenant, conformément à la démarche générale, nous introduisons l'indice brut

$$s(\omega, \pi) = \text{card}[\text{gr}(\omega) \cap (S(\pi) \times R(\pi))], \quad (18)$$

auquel nous associons l'indice aléatoire

$$s(\omega, \pi^*) = \text{card}[\text{gr}(\omega) \cap (S(\pi^*) \times R(\pi^*))], \quad (19)$$

où π^* est une partition aléatoire dans l'ensemble $\mathfrak{T}\ [n;t\ (\pi)]$ des partitions de même type que celui de π, muni d'une probabilité uniformément répartie.

Le calcul de la moyenne et de la variance de $s(\omega,\pi^*)$ s'effectue de façon exacte. D'autre part, cette v.a. est — sous des conditions assez générales —asymptotiquement normale [Lerman (1976), (1981) Chap. 4, (1983)].

L'indice $s\ (\omega,\pi)$ centré et réduit :

$$\sum(\omega, \pi) = [s(\omega, \pi) - \mathcal{E}\,(s\,(\omega, \pi^*))]/\sqrt{\text{var}(s(\omega, \pi^*))} \quad (20)$$

représente un critère d'évaluation de l'adéquation d'une partition π à l'ordonnance ω qui constitue l'information ordinale résultant des ressemblances mutuelles —mesurées par un indice — entre éléments de l'ensemble à classifier.

$\Sigma(\omega,\pi)$ prend le nom de « Statistique *globale* des niveaux » lorsqu'on considère la suite de ses valeurs sur la suite des niveaux de l'arbre des classifications :

$$\left\{\sum(\omega, \pi_i)/1 \leqslant i \leqslant k\right\}. \quad (21)$$

C'est l'examen de cette suite observée de valeurs (cf. Figure 4) qui permet de détecter les niveaux les plus « significatifs » qui correspondent aux principaux états d'équilibre dans la synthèse classificatoire que fournit l'arbre. Ces niveaux correspondent aux maxima locaux les plus nets de la fonction (21).

Pour recouvrir la notion de « nœud significatif » qui ponctue l'achèvement d'une sous-classe ou d'une classe, on pourra partir de l'indice brut « local » :

$$s(\omega, i) = \text{card}[\text{gr}(\omega) \cap (S(\pi_i) \times R_1(\pi_i))], \quad (22)$$

où $R_1(\pi_i)$ est l'ensemble des paires réunies pour la première fois au niveau i de l'arbre par la partition π_i.$R_1(\pi_i)$ est un ensemble de paires dont les deux composantes appartiennent respectivement à deux sous classes.

L'indice aléatoire que nous pouvons associer à $s(\omega,i)$ est

$$s(\omega^*, i) = \text{card}[\text{gr}(\omega^*) \cap (S(\pi_i) \times R_1(\pi_i))], \quad (23)$$

où ω^* est une préordonnance aléatoire dans l'ensemble — muni d'une probabilité uniforme — de toutes les préordonnances possibles de même type cardinal. Il en résulte une première « statistique locale des niveaux » :

$$\theta(\omega, i) = [s(\omega, i) - \mathcal{E}\,(s(\omega^*, i))]/\sqrt{\text{var}(s(\omega, \pi^*))}, \quad (24)$$

Une deuxième « statistique locale des niveaux » peut directement être définie comme le taux d'accroissement de la statistique globale entre deux niveaux consécutifs :

$$\tau(\omega, i) = \left[\sum(\omega, \pi_i) - \sum(\omega, \pi_{(i-1)})\right]. \quad (25)$$

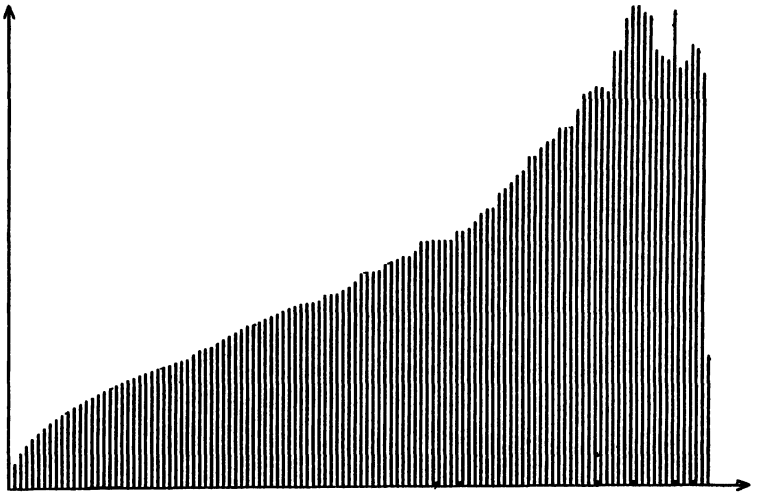

Figure 4 : Évolution de la « Statistique globale » sur la suite des niveaux dans un cas réel (l'axe horizontal indique la suite des niveaux et l'axe vertical, la valeur de la statistique).

Certes, une certaine différence d'appréciation peut apparaître entre les deux formes de la statistique locale des niveaux (24) ou (25); mais, la tendance générale est à la croissance lorsqu'une classe en cours de formation se confirme, à la décroissance devant l'arrêt de l'élaboration d'une classe ayant une certaine consistance, au profit de l'accroissement d'autres classes plus embryonnaires. Un « nœud significatif » correspond dans ces conditions à un maximum local de la suite des valeurs $\{\theta(\omega,i)/1 \leqslant i \leqslant k\}$ (resp. $\tau(\omega,i)/1 \leqslant i \leqslant k\}$). La représentation de l'arbre que nous fournissons dans notre méthode, est — comme nous l'avons déjà mentionné ci-dessus — condensée aux niveaux où apparaît un nœud significatif qui est marqué en tant que tel par une étoile (cf. Figures 5 et 6).

L'importance des valeurs atteintes par $\Sigma(\omega,i)$ [resp. $\theta(\omega,i)$] a certes un intérêt; mais, il faut savoir que ce qui importe pour l'interprétation dynamique de l'arbre des classifications, c'est bien davantage l'évolution de la suite des valeurs de Σ (resp. θ).

Ce souci de reconnaître des composants pertinents et bien marqués dans la construction d'un arbre des classifications, existe bien dans d'autres approches métriques (M. Jambu et M.O. Lebeaux (1978)]. Cependant, ces dernières utilisent les mêmes critères que ceux qui ont prévalu à la formation des classes. Or, il y a de « bons » critères pour la formation des classifications et de « bons » critères pour leur évaluation, *mais ce ne sont pas nécessairement les mêmes.*
Nous préférons quant à nous considérer un critère très général où on ne retient que l'aspect ordinal des structures de ressemblance à comparer.

Un autre aspect d'une validation intrinsèque des résultats concerne la détermination des éléments plus ou moins « moteurs » (ou à l'opposé « neutres ») dans l'entraînement d'une classe. A cet égard, nous proposons de mesurer le degré de neutralité d'un élément α par la petitesse de la variance de ses proximités aux autres éléments de l'ensemble à organiser. Pour fixer les idées, en reprenant le cas où cet ensemble est une famille $\mathcal{A}$ d'attributs, cet indice peut prendre la forme suivante

$$\mathfrak{D}(\alpha) = \frac{1}{(p-1)} \sum \{[Q_s(\alpha, b) - Q_s(\alpha)]^2 / b \in - \{\alpha\}\}$$

où $p = \text{card}(\mathcal{A})$ et où

$$Q_s(\alpha) = \frac{1}{(p-1)} \sum \{Q_s(\alpha, b) / b \in \mathcal{A} - \{\alpha\}\},$$

$Q_s(a,b)$ ayant été défini au paragraphe précédent (cf. formule 9*s*).

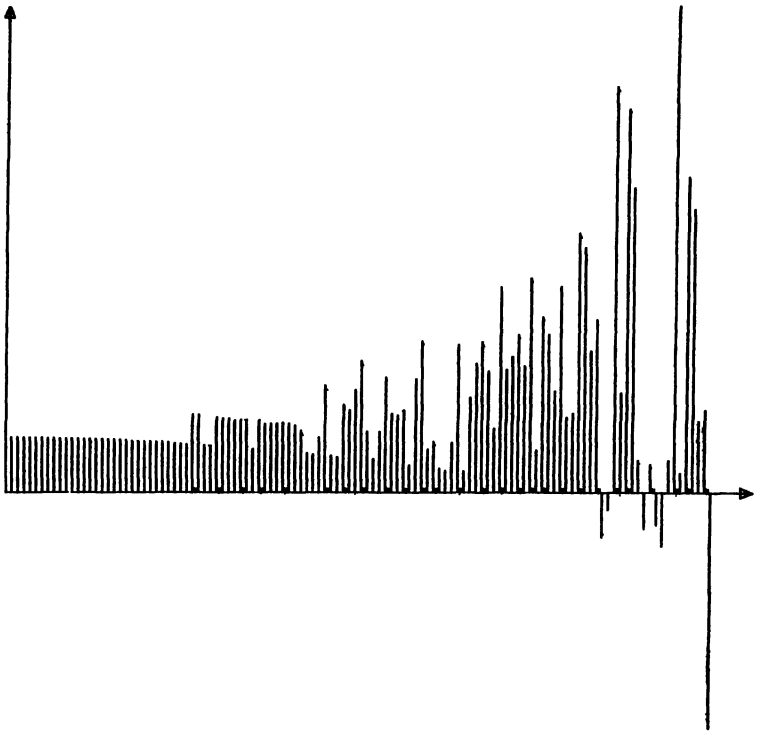

Figure 5 : Évolution de la « Statistique locale » sur la suite des niveaux dans un cas réel.

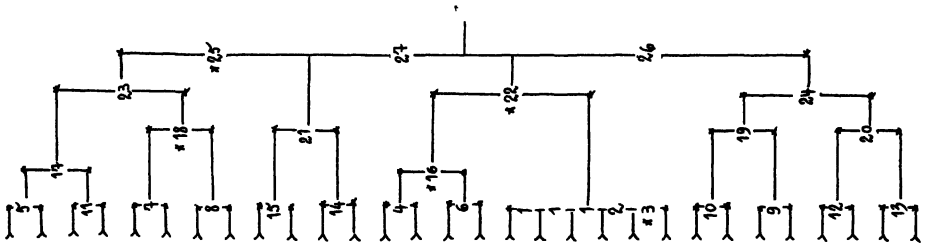

Figure 6 : Exemple de la représentation d'un arbre de classification condensé aux niveau où apparaît un nœud significatif marqué par une étoile (*). Les niveaux de formation des autres nœuds sont également marqués. Les libellés des feuilles de l'arbre ont été expressément enlevés.

Pour nous résumer, nous avons commencé par justifier au mieux notre façon d'évaluer les ressemblances entre parties disjointes de l'ensemble à classifier. Dans cette approche, les hypothèses statistiques d'absence de liaison interviennent de façon cruciale. Indépendamment et supposant acquis la donnée d'un indice d'association ou de similarité (resp. dissimilarité) entre éléments, la forme de *validation* dont il vient d'être question ci-dessus, cherche à évaluer à partir de critères les plus généraux et exogènes, ce qui se distingue dans la formation de l'arbre des classifications. Par ailleurs, la conception de ces critères se fait conformément à une démarche générale que nous avons schématisée au paragraphe précédent (cf. (7) § II).

Une deuxième forme ultime de validation « oublie » l'indice d'association ou de similarité (resp. dissimilarité) qui a permis l'émergence d'une classification qui a été retenue et qu'on suppose acquise. Une telle classification peut par exemple correspondre à un niveau hautement significatif de l'arbre des classifications. Cette forme de validation qui correspond à l'interprétation et l'« explication » des

résultats, se fait par un retour et une réorganisation du tableau des données brutes. On peut à ce niveau opérer des croisements de classifications :

— Croisement entre deux classifications sur l'ensemble des objets (une même classification résulte de la description par un même ensemble de variables).

— Croisement entre une classification sur l'ensemble des objets (aux classes non ordonnées ou ordonnées) et une classification sur l'ensemble des variables (aux classes non ordonnées ou ordonnées).

Un problème d'indices se pose pour charger de façon adéquate les cases du croisement, compte tenu de la nature des variables descriptives.

On peut également à ce niveau définir le « degré de responsabilité » d'une variable ou d'une classe de variables, dans la formation d'une classe d'objets (resp. d'un objet ou d'une classe d'objets, dans la formation d'une classe de variables) (cf. [3], [15], [16] et [21]).

Nous avons pu nous rendre compte ci-dessus comment un développement original de la Statistique non paramétrique pouvait conduire à l'élaboration des indices et critères qui permettent de bâtir et d'évaluer les résultats d'une classification automatique des données. Cette approche ne fait nullement référence dans ses aspects statistiques à des hypothèses formulées au niveau d'une hypothétique population mère $\mathfrak{T}$.

Mais, pour étudier les problèmes de stabilité évoquées dans l'introduction, il importe de considérer de telles hypothèses.

IV — Aspect stabilité

Reprenons ici le contexte du paragraphe I ci-dessus. E : ensemble des objets est regardé comme la réalisation d'un échantillon aléatoire de taille n provenant d'une population mère $\mathfrak{T}$ de taille N. La donnée est une famille $\mathfrak{A} = \{a_j / j \in J\}$ d'attributs de description qu'il s'agit d'organiser en classes et sous-classes d'association à partir de leur observation sur E.

On suppose que sur $\mathfrak{T}$, se trouve donnée la distribution jointe de $\mathfrak{A}$. Relativement à un même couple d'attributs logiques (a,b), reprenons les notations du paragraphe I :

$\pi(a), \pi(b), \pi(a \wedge b) \longrightarrow \rho\,(a,b)$ au niveau de $\mathfrak{T}$

$p\,(a), p(b), p(a \wedge b) \longrightarrow r(a,b)$ au niveau de E

$P\,(a), P(b), P(a \wedge b) \longrightarrow R(a,b)$ au niveau de $\mathcal{E}$,

où $\rho(a,b)$ [resp. $r\,(a,b)$ et $R(a,b)$] représente le coefficient de K. Pearson défini au niveau de $\mathfrak{T}$ (resp. au niveau de E et de l'échantillon aléatoire $\mathcal{E}$).

On démontre (cf. [14]) que $R(a,b)$ suit asymptotiquement une loi normale de moyenne $\rho(a,b)$ et de variance C^2/n. L'approximation normale est en général excellente. D'autre part, on peut se rendre compte que même dans le cas de forte dépendance — définie au niveau $\mathfrak{T}$ — on a $0{,}5 < C^2 < 1{,}5$. De toute façon, un théorème assure que $C^2 = 1$ en cas d'indépendance et l'expérience montre que pourvu que la dépendance ne soit pas trop forte, C^2 reste au voisinage de 1.

Ce type de résultats concernant la comparaison de deux variables seulement est dans la lignée de recherches initialisées par L.A. Goodman et W.H. Kruskal (cf. [4] et [5]) qui se sont intéressés au cas de la comparaison de deux variables qualitatives nominales ou ordinales, mais qui, curieusement, n'avaient pas pris en compte le cas de la comparaison de deux attributs logiques qui s'est avéré très riche (cf. [14]) et pour lequel l'expression formelle des indices tel que celui de K. Pearson, s'impose. En effet, les expressions formelles des indices proposés par ces auteurs n'est pas sans un certain arbitraire puisqu'elles résultent de la simple intuition; alors que pour nous, comme nous l'avons souligné (cf. (7) § II), cette expression formelle est construite à partir d'une hypothèse adéquate d'absence de liaison. D'ailleurs, le coefficient γ de Kruskal pour la comparaison de deux variables qualitatives ordinales s'avère « biaisé », en ce sens que l'espérance de son numérateur n'est pas nulle.

Sur la base de ces résultats statistiques, l'intervalle de confiance symétrique pour $\rho(a,b)$, au seuil $(1\text{-}\alpha)$, prend la forme suivante :

$$\left[r(a,b) - \sqrt{C^2/n}\,\Phi^{-1}\left(1 - \frac{\alpha}{2}\right),\ r(a,b) + \sqrt{C^2/n}\,\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right]. \tag{1}$$

Dans ces conditions, pour un lien relativement « faible » — comme c'est le cas fréquent en analyse des données — et un seuil de confiance de 0,99, l'intervalle de confiance devient

$$[r(a, b) - 2.5/\sqrt{n},\ r(a, b) + 2.5/\sqrt{n}]. \quad (2)$$

Il s'avère que même pour $n = 10^4$, il y a — en termes d'intervalle de confiance — un doute après le *premier* chiffre significatif!

C'est peut-être ce fait qui explique l'attitude de relative méfiance des Statisticiens classiques vis-à-vis de l'Analyse des Données. Mais ces derniers oublient que le problème n'est pas tant d'estimer une corrélation ou des corrélations, il consiste à *organiser* les éléments conformément aux corrélations. Pour illustrer ce point, considérons un couple de paires d'attributs $(\{a,b\},\{c,d\})$ où $\{a,b\}$ et $\{c,d\}$ sont sans composante commune. Imaginons que $\rho\,(a,b) = 0$ et $\rho\,(c,d) = 0{,}08$ [faibles dépendances entre a et b d'une part, c et d d'autre part, mais tout de même $\rho\,(a,b) < \rho\,(c,d)$]. Dans ces conditions, $[R(a,b) - R(c,d)]$ suit — asymptotiquement — une loi normale de moyenne $[\rho\,(a,b - \rho\,(c,d)]$. Toujours pour $n = 10^4$, on a :

$$Pr\{R(a, b) < R(c, d)\} = \Phi(8/\sqrt{2}) \cong 1. \quad (3)$$

D'où la très grande stabilité dans la préservation de l'inégalité $\rho\,(a,b) < \rho\,(c,d)$.

Toutefois, le résultat précédent (cf. (2)) montre toute l'aberration de la recherche d'une grande précision de calcul des coefficients de corrélation entre variables pour une analyse des données. En pondérant par une indication telle que (3), on peut suggérer une précision calcul où le nombre de chiffres significatifs serait de l'ordre de $\mathrm{Log}_{10} n$.

Nous allons à présent reprendre le problème de la distribution de la table des indices aléatoires d'association, mais par rapport à l'optique inférentielle considérée ici, où on suppose l'indépendance au niveau de $\mathfrak{T}$, entre les différents attributs $a_j, 1 \leqslant j \leqslant p$, (cf. [14] pour un développement détaillé).

L'observation d'un individu correspond à la réalisation jointe $(\varepsilon_1, \varepsilon_2, ..., \varepsilon_j, ..., \varepsilon_p)$ de la suite $(a_1, a_2, ..., a_j, ..., a_p)$ d'attributs logiques. Le vecteur de Booléens $(\varepsilon_1, \varepsilon_2, ..., \varepsilon_j, ..., \varepsilon_p)$ appartient au cube $\{0,1\}^p$ catégories, avec les probabilités.

$$\pi(\varepsilon_1, \varepsilon_2, ..., \varepsilon_j, ..., \varepsilon_p) = \prod_{1 \leqslant j \leqslant p} \pi(\varepsilon_j), \quad (4)$$

où $\pi(\varepsilon_j)$ est une probabilité définie au niveau de $\mathfrak{T}$ par la proportion de sujets pour lesquels la valeur de a_j est ε_j.

L'observation de l'échantillon E charge les cellules de $\{0,1\}^p$ au moyen de

$$\{n(\varepsilon_1, \varepsilon_2, ..., \varepsilon_p)/(\varepsilon_1, \varepsilon_2, ..., \varepsilon_p) \in \{0,1\}^p\}. \quad (5)$$

$\mathcal{E}$ charge $\{0,1\}^p$ du vecteur multinomial

$$\{n^*(\varepsilon_1, \varepsilon_2, ..., \varepsilon_p)/(\varepsilon_1, \varepsilon_2, ..., \varepsilon_p) \in \{0,1\}^p\} \quad (6)$$

à 2^p composantes, de paramètres n et les probabilités (4). A cet égard, on connaît bien l'approximation normale (ici à $(2^p - 1)$ dimensions) de la loi multinomiale (ici à 2^p catégories).

L'indice « brut » entre deux attributs a_j et a_k se met sous la forme d'une somme à $2^{(p-2)}$ termes :

$$n(a_j \wedge a_k) = \sum \{n(\varepsilon_1, \varepsilon_2, ..., \varepsilon_p)/\varepsilon_j = \varepsilon_k = 1\}. \quad (7)$$

L'indice « brut » aléatoire associé se met sous la forme

$$n^*(a_j \wedge a_k) = \sum \{n^*(\varepsilon_1, \varepsilon_2, ..., \varepsilon_p)/\varepsilon_j = \varepsilon_k = 1\}. \quad (8)$$

Compte tenu de cette linéarité de l'expression, il en résulte que la loi jointe des $p\,(p-1)/2$ *v.a.* :

$$\{n^*(a_j \wedge a_k)/1 \leqslant j < k \leqslant p\}, \quad (9)$$

est asymptotiquement multinormale. La matrice de covariance V peut être calculée en fonction de n et des probabilités $\pi(1_j)$, $1 \leqslant j \leqslant p$.

Si on considère à présent le vecteur (9) des *v.a.* centrées :

$$\{q^*(a_j, a_k) = [n^*(a_j \wedge a_k) - n\pi(1_j)\pi(1_k)]/1 \leqslant j < k \leqslant p\}, \quad (10)$$

la statistique ${}^t q^* V_q^*$ — où ^{t} désigne la transposition — suit une loi de χ^2 à $\binom{p}{2}$ degrés de liberté.

Il apparaît ainsi, dans ce contexte inférentiel une nouvelle forme de réduction globale des similarités ou associations — au moyen de $\left[{}^t q V q \Big/ \binom{p}{2}\right]^{1/2}$ — avant la référence à une échelle de probabilité pour la mesure des liaisons définie par la loi normale.

Plus précisément, en estimant $\pi\,(a_j)$ par la proportion $p\,(a_j)$ observée au niveau de l'échantillon E, $1 \leqslant j \leqslant k$, on considérera la table des indices

$$\{Q_s(a_j, a_k) = \frac{n(a_j \wedge a_k) - np(a_j)p(a_k)}{\sqrt{{}^t q V q \Big/ \binom{p}{2}}} \,/\, 1 \leqslant j < k \leqslant p,\} \quad (11)$$

où q désigne le vecteur colonne

$${}^t\{[n(a_j \wedge a_k) - np(a_j)p(a_k)]/1 \leqslant j < k \leqslant p\}$$

L'indice « vraisemblance du lien », relativement à cette nouvelle forme de la modélisation de l'hypothèse d'absence de liaison, se met sous la forme

$$P_s(a_j, a_k) = \Phi[Q_s(a_j, a_k)], \quad (12)$$

$1 \leqslant j \leqslant k \leqslant p$.

L'étude de la distribution de la suite de *v.a.* (9) ci-dessus ou (10) du paragraphe II, peut être considérée comme un correspondant non paramétrique de l'étude classique de Wishart (cf. [23]).

RÉFÉRENCES

[1] ALLAIS M. — Fréquence, probabilité et hasard. *Journal de la Société de Statistique de Paris*, n° 2, 2e trimestre (1983).

[2] BLANCARD M. — Analyse d'un important fichier de bilans de santé — Rapport de DEA, Univ. de Rennes I, sept. (1976).

[3] GEFFRAULT J.P. — Discrimination de classes et détermination d'ensembles minimaux de mesures pour la classification automatique de formes. Application à des données en Archéologie — Thèse de 3e cycle, Univ. de Rennes I, mars 1982.

[4] GOODMAN L.A. and KRUSKAL W.H. — Measures of association for cross classification, approximate sampling theory — J.A.S.A. 58, June (1963), 310-364.

[5] GOODMAN L.A. and KRUSKAL W.H. — Measures of association for cross classification, IV : simplification of asymptotic variances — J.A.S.A. 67, June (1972), 415-421.

[6] JAMBU M. et LEBEAUX M.O. — Classification automatique pour l'analyse des données — Dunod, Paris (1978).

[7] KENDALL M.G. — Rank correlation methods — Charles Griffin, fourth edition, London (1970).
[8] LERMAN I.C. — Les bases de la classification automatique — Gauthier-Villars, Paris (1970a).
[9] LERMAN I.C. — Sur l'analyse des données préalable à une classification automatique. Proposition d'une nouvelle mesure de similarité — Rev. Math. et Sc. Hum., 8e année, nº 32, (1970b).
[10] LERMAN I.C. — Étude distributionnelle de statistiques de proximité entre structures finies de même type; Application à la classification automatique — Cahiers du B.U.R.O. nº 19, Paris (1973), 1-52.
[11] LERMAN I.C. — Formal analysis of a general notion of proximity between variables — In Proceedings « Congrès Européen des Statisticiens, Grenoble », sept. 1976, published by North Holland in 1977.
[12] LERMAN I.C. — Classification et analyse ordinale des données — Dunod, Paris (1981).
[13] LERMAN I.C. — Sur la signification des classes issues d'une classification automatique des données — NATO ASI Series, Vol. G1, Numerical Taxonomy. Edited by J. Felsenstein, Springer Verlag, (1983), 179-198.
[14] LERMAN I.C. — Justification et validité statistique d'une échelle [0,1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées — Publications Inst. Stat. de l'Univ. de Paris, Vol. XXIX, fasc. 3-4, Paris (1984), 27-59.
[15] LERMAN I.C., HARDOUIN M. et CHANTREL T. — Analyse de la situation relative entre classifications « floues » — 2e Journées Analyse des Données et Informatique, Versailles 1979, Data Analysis and Informatics, E. Diday et al. eds, North Holland, (1980).
[16] LERMAN I.C. et Collaborateurs — Programmes d'analyse des résultats d'une classification automatique — Publication Interne IRISA nº 178, sept. (1982), 79 pages.
[17] LERMAN I.C. et PETER Ph. — Élaboration et logiciel d'un indice de similarité entre objets d'un type quelconque — Publ. Int. IRISA nº 262, juillet (1985), 72 pages.
[18] NICOLAÜ M.H. — Analyse d'un algorithme de classification — Thèse de 3e cycle, Univ. Paris VI, ISUP, (1972).
[19] NICOLAÜ M.H. — Contribuiçoes ao estudo dos coeficientes de comparação em análise classificatória — Thèse de doctorat, Fac. des Sc. de Lisbonne, févr. (1981).
[20] NICOLAÜ F. — Criterios de analise classificatoria hierarquica baseados ma funçao de distribuiçao — Thèse de Doctorat, Faculté des Sciences de Lisbonne, février (1981).
[21] PROD'HOMME A. — Indice d'explication des classes obtenues par une méthode de classification hiérarchique respectant la contrainte de contiguïté spatiale. Application à la viticulture Girondine et à la construction de logements dans les Bouches-du-Rhône — Thèse de 3e cycle, Univ. de Rennes I, décembre 1980.
[22] WALD A. and WOLFOWITZ J. — Statistical tests based on permutations of the observations — Ann. Math. Stat. 15, (1944), 358-372.
[23] WISHART J. — The generalized product moment distribution in samples from a normal multivariate population — Biometrika, Vol. 20A, (1928), 32-52.