Classification and Related Methods of Data Analysis H.H. Bock (Editor) North Holland, 1988.

COMPARING PARTITIONS (MATHEMATICAL AND STATISTICAL ASPECTS)

I.C. LERMAN

IRISA - Campus de Beaulieu, 35042 Rennes Cédex, France.

The formal analysis of the most of the comparison coefficients between two partitions, shows the non intervention of the relational constraint which results from the partition structure. This constraint must appear in the standardization of the considered coefficient ; that is to say, at the level of the coefficient denominator. This standardization has 'formal' or 'statistical' nature. For the first, we show how -by replacing the notion of mathematical formulae by that of recursive algorithm- we resolve a combinatorial optimization problem which has been considered as yet as very difficult. On the other hand, we study the asymptotic formal expression of the coefficient obtained by statistical standardization with respect to an adequate hypothesis of no relation. This last coefficient takes closely into account the relational constraint.

1. INTRODUCTION

From the beginning of our research, the point of view that we have adopted and developed consists of considering a descriptive qualitative variable of a set O of objects as defining a relation on the set O. According to its complexity this relation can be represented by a subset of O or of OxO or of (OxO)x(OxO) [Lerman(1970), (1973),(1981),(1987)]. A qualitative nominal variable c is a particular case of relational qualitative variable which induces a partition $\pi=\{A_i/1\leqslant i\leqslant I\}$ into I non empty classes. We may represent such partition at the level of the set $O^{\{2\}}=P_2(O)$ of unordered object pairs, namely, by the subset $R(\pi)$ of the object pairs that π joins, more precisely

$$R(\pi)=\sum\{A_i^{\{2\}}=P_2(A_i)/1\leqslant i\leqslant I\} \quad \text{(set sum)}$$

We may also introduce the subset $S(\pi)$ of object pairs separated by the partition : $S(\pi)=\sum\{A_i*A_{i'}/1\leqslant i<i'\leqslant I\}$, where $A_i*A_{i'}=\{\{x,y\}/x\in A_i,\ y\in A_{i'}\}$. $R(\pi)$ and $S(\pi)$ determine a partition with two classes of $O^{\{2\}}$.

Thus, comparing two partitions π and $\times$ can be expressed in terms of the comparison of subsets of the set $O^{\{2\}}$ of unordered distinct object pairs. Indeed we can also associate to $\times=\{B_j/1\leqslant j\leqslant J\}$, the representation set $R(\times)$ and $S(\times)$.

Most of the coefficients proposed in the literature use the following raw indices : $s=\text{card}[R(\pi)\cap R(\times)]$, $u=\text{card}[R(\pi)\cap R^c(\times)]$, $v=\text{card}[R^c(\pi)\cap R(\times)]$, $t=\text{card}[R^c(\pi)\cap R^c(\times)]$ where $R^c(\pi)$ [resp. $R^c(\times)$] denotes the complementary subset of $R(\pi)$ [resp.$R(\times)$] in $O^{\{2\}}$. For example :

Rand(1971): $[(s+t)/(s+u+v+t)]$. (1)
Jaccard(1908): $[s/(s+u+v)]$. (2)
Fowlkes and Mallows (1983): $\{s/\sqrt{[(s+u)(s+v)]}\}$. (3)

The coefficient (2) is correctly attributed to Jaccard. But it is not exactly the case for the two other indices that had been considered a long time before, but in a very different context. Around 1950 to 1960, several similarity indiceshad been considered for comparing individuals described by 0-1 logical attributes. If C denotes the featu-

re set individual x can be represented by the subset C_x ($C_x \subset C$) of the features that he has got. Comparing two individuals x and y is then equivalent to comparing the two associated subsets C_x and C_y of C, from the parameters :s=card($C_x \cap C_y$), u=card($C_x \cap C_y^c$), v=card($C_x^c \cap C_y$), t=card($C_x^c \cap C_y^c$) where C_x^c (resp. C_y^c) indicates the complementary subset of C_x (resp.C_y) in C.

Then we can realize immediatly that the Rand coefficient is nothing else than that of Sokal and Michener (1958) which can also be written as $\{1-[(u+v)/c]\}$, where c=card(C)=s+u+v+t. On the other hand Fowlkes and Mallows' coefficient is nothing but the coefficient of Ochiai (1957). Furthermore, a coefficient of 'Goodman and Kruskal' type (1954): $\{[(s+t)-(u+v)]/(s+t+u+v)\}$ may be written $\{1-[2(u+v)/c]\}$ and it corresponds exactly to the Hamann index (1961). Now, if we look at the object x (resp. y) as represented by a total preorder with two classes C_x^c and C_x, where $C_x^c < C_x$ (resp. C_y^c and C_y where $C_y^c < C_y$), the Goodman and Kruskal coefficient is exactly the Yule coefficient [(1911),(1912)]: $[(st-uv)/(st+uv)]$.

Therefore, if we reduce the comparison of two partitions π and χ, to the comparison of the two subsets $R(\pi)$ and $R(\chi)$ of the finite set $O^{\{2\}}$ -without going more deeply about the nature of the structures to be compared- we may consider each of the similarity indices considered by the taxonomists to compare taxons described by 0-1 logical attributes. Afterwards, we may explicit a given similarity index with respect to the contingency table $\{c_{ij}/1\leq i\leq I,\ 1\leq j\leq J\}$ that crosses the two partitions π and χ: c_{ij}=card($A_i \cap B_j$), a_i=card(A_i), b_i=card(B_i), $1\leq i\leq I$, $1\leq j\leq J$. In fact we have :

$\text{card}[O^{\{2\}}]=\binom{n}{2}$, where n is the cardinal of O,

$\text{card}[R(\pi)]=\sum\{\binom{a_i}{2}/1\leq i\leq I\}$, $\text{card}[R(\chi)]=\sum\{\binom{b_j}{2}/1\leq j\leq J\}$

$s=\text{card}[R(\pi)\cap R(\chi)]=\sum\{\binom{c_{ij}}{2}/1\leq i\leq I,\ 1\leq j\leq J\}$

$u=\text{card}[R(\pi)\cap S(\chi)]=\sum\{c_{ij}c_{ij'}/1\leq i\leq I,\ 1\leq j<j'\leq J\}$

$v=\text{card}[S(\pi)\cap R(\chi)]=\sum\{c_{ij}c_{i'j}/1\leq i<i'\leq I,\ 1\leq j\leq J\}$

$t=\text{card}[S(\pi)\cap S(\chi)]=\sum\{c_{ij}c_{i'j'}/1\leq i<i'\leq I,\ 1\leq j<j'\leq J\}$,

where $S(\pi)=R^c(\pi)$ [resp.$S(\chi)=R^c(\chi)$].

All these coefficients -that we obtain by analogy with known similarity indices between subsets of a finite set- do not take completely into account the specific nature of the partition structures to be compared. As a matter of fact, we may consider these indices to compare any two binary relations that we represent by their respective graphs at the level of OxO. But the representation set has particular meaning in comparing partitions, since it consists of element pairs. Though the subsets $R(\Pi)$ and $R(\chi)$ are closed transitively, the transitivity constraint does not intervene in the comparison, for the previously mentioned indices. F. Marcotorchino (1984) showed that several association coefficients between partitions -proposed in the literature- correspond to comparing two 'linear' codings of the couple set OxO. These last are not specific of comparing two partitions. Formally they can be considered to compare any two binary relations, or even two subsets of a finite set.

2. GENERAL DIAGRAM FOR COMPARING TWO RELATIONAL VARIABLES

To compare two relational qualitative variables, we have in our research set up the following diagram :

$$(\alpha,\beta)\in \Omega_\alpha \times \Omega_\beta \longrightarrow [R(\alpha),R(\beta)]\in \Omega_\alpha \times \Omega_\beta$$

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

- Hypothesis of non link (h.n.l.) (or independence) taking into account -in a strict or fuzzy ways- the cardinal characteristics of α and β.

$$\longrightarrow S=s(\alpha^*,\beta^*)=\mathrm{card}\left[R(\alpha^*)\cap R(\beta^*)\right]$$

$$\longrightarrow Q(\alpha,\beta)=\left[s-\mathscr{E}(S)\right]/\sqrt{\left[\mathrm{var}(S)\right]}.$$

In this diagram α and β are the two relations on the set O of objects, respectively determined by the two variables to be compared. $\mathcal{A}$ (resp. $\mathcal{B}$) is the set of all relations of the "same type" as α (resp. β). $R(\alpha)$ [resp. $R(\beta)$] is the set representation of α (resp. β). $R(\alpha)$ [resp. $R(\beta)$] is a subset of O, or of OxO, or even of (OxO)x(OxO). Ω_α (resp. Ω_β) is the set of all possible representation subsets of a relation of the same type as α (resp. β). $s=s(\alpha,\beta)$ is called the 'raw' index. α^* and β^* are two independent relational random variables, respectively associated to α and β, according to a hypothesis of no relation (h.n.l.) which takes into account -in a strict of fuzzy way- the cardinal characteristics of α and β. S is the "random raw index", the mathematical expectation and variance of which being denoted by $\mathscr{E}(S)$ and var(S). $Q(\alpha,\beta)$ is the "standardized index".

We have extensively used this previous diagram in the elaboration of our total or partial association coefficients between qualitative variables [Lerman (1973),(1981), (1983a),(1983b)]. To make this diagram completely clear, let us illustrate it in the case of concern ; that is to say, the comparison between two partitions.

α and β are two partitions with -without loss of generality- labelled classes. According to above, we will denote then by π and $\times$. $t(\pi)$ [resp.$t(\times)$] indicates the 'type' of the partition π(resp.$\times$) ; that is to say, the ordered sequence of the class cardinals : $t(\pi)=(a_i/1\leqslant i\leqslant I)$ [resp. $t(\times)=(b_j/1\leqslant j\leqslant J)$. In these conditions, $\mathcal{A}$ (resp.$\mathcal{B}$) is the set of labelled partitions on O, whose type is $t(\pi)$ [resp.$t(\times)$]. $R(\pi)$ [resp.$R(\times)$] is the set of object pairs, such as the two components are joined in a same class of π (resp. $\times$) (cf. above). Ω_π (resp. $\Omega_\times$) may be defined as the subset set of $O^{\{2\}}$, each element (subset of $O^{\{2\}}$) of which, corresponding to the representation of a partition of type $t(\pi)$ [resp.$t(\times)$]. We have yet expressed $s=\mathrm{card}\left[R(\pi)\cap R(\times)\right]$ [cf. (4) above].

There are three fundamental forms of the 'h.n.l.' [Lerman(1981) Chap.2]. We consider here the strict form where π^*(resp.$\times^*$) is a random partition in the set $\mathcal{A}$ (resp. $\mathcal{B}$) provided by an uniform probability measure. The mean and variance of the random raw index $S=s(\pi^*,\times^*)$ are given by [Lerman(1973),(1981)] :

$\mathscr{E}(S)=\lambda\mu$ and $\mathrm{var}(S)=\lambda\mu+\rho\sigma+\theta\zeta-\lambda^2\mu^2$

where

$$\lambda=\sum\{a_i(a_i-1)/\sqrt{[2n(n-1)]}\mid 1\leqslant i\leqslant I\} \qquad (1)$$

$$\rho=\sum\{a_i(a_i-1)(a_i-2)/\sqrt{[n(n-1)(n-2)]}\mid 1\leqslant i\leqslant I\}$$

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

and where the expressions of μ,σ,ζ have respectively got the same forms as λ,ρ,θ ; the a_i of $t(\pi)$ being replaced by the b_j of $t(\times)$.

We may notice that θ can be expressed in terms of λ and ρ. Let us situate two classical association coefficients with respect to the preceding diagram. The first -of comparing two logical attributes a and b -is that of K. Pearson (1928) and the second- of comparing two ranking variables r and s- is that of M.G. Kendall(1970).

In order to obtain K. Pearson's coefficient, we represent a logical attribute a (resp. b) by the subset O(a) [resp.O(b)] of the objects which possesse the feature a (resp. b). Therefore R(a)=O(a) [resp.R(b)=O(b)]. The 'h.n.l.' is strict and it associates to R(a) [resp.R(b)], a random subset R(a*) [resp.R(b*)] in the set -provided by an uniform probability measure- of all subsets of O, with the same cardinal n(a) [resp. n(b)]. The random variable S is hypergeometric. Q(a,b) can be written as follows :

$$Q(a,b)=\sqrt{(n-1)}\ \rho(a,b)\simeq\sqrt{n}\ \rho(a,b). \qquad (2)$$

where $\rho(a,b)$ is a pure coefficient, ranging from -1 to 1, the limit of which -when n tends to infinity and when $[n(a)/n]$ (resp. $[n(b)/n]$) tends to a limit different from zero or one- being independent of n.

We may notice that the (a,b) coefficient can be obtained by means of one of the two following expressions :

$$(i)\ \rho(a,b) = \frac{1}{\sqrt{n}}\ Q(a,b) \qquad (3)$$

$$(ii)\ \rho(a,b) = \frac{Q(a,b)}{\sqrt{Q(a,a)Q(b,b)}} \qquad (4)$$

Let us now consider how to obtain -in the framework of the diagram- the M.G. Kendall τ coefficient of comparing two ranking variables r and s. Each variable defines a total and strict order on the object set O. R(r) [resp.R(s)] is the graph in OxO of the total order relation defined by r (resp.s) that we denote also by r (resp. s). r* (resp.s*) is a random total order in the set -provided by an uniform probability measure- of the n! strict and total orders on O. In these conditions, the M.G.Kendall τ coefficient can be put in the following form :

$$\tau(r,s) = \frac{\{\, s(r,s) - \mathcal{E}[s(r^*,s^*)] \,\}}{\{\max[s(r',s')] - \mathcal{E}[s(r^*,s^*)] \}} \qquad (5)$$

where $\max[s(r',s')]$ is the maximum possible of the raw index s(r,s), it concerns n(n-1)/2.

Some researchers like L. Hubert and P. Arabie (1985) consider association coefficient between two qualitative variables α and β as necessarily having the same form as (5). Namely and in the most general case

$$\tau(\alpha,\beta) = \frac{\{s(\alpha,\beta) - \mathcal{E}[s(\alpha^*,\beta^*)]}{\{\max[s(\alpha',\beta')] - \mathcal{E}[s(\alpha^*,\beta^*)]\}} \qquad (6)$$

where $\max[s(\alpha',\beta')]$ is the maximum possible of $\mathrm{card}[R(\alpha') \cap R(\beta')]$ for α' (resp. β') of the same type as α (resp. β).

On the other hand, they consider that the statistic $Q(\alpha,\beta)$ should be devoted to testing independence hypothesis between α and β. But we showed the non relevance of the tests of independence hypothesis between descriptive variables in Data Analysis [Lerman(1984)]. Then nothing forbids to consider a coefficient $\rho(\alpha,\beta)$ directly deduced from $Q(\alpha,\beta)$ in the same way as the $\rho(a,b)$ K. Pearson coefficient can be deduced from Q(a,b) [cf.(3) and (4)].

Our goal in the following is to give a brief look on the most recent results that we have obtained on the possibility and the analysis of formulae such as (3),(4) and specially (5) in case of comparing two partitions π and $\times$. On the contrary of the coefficients presented at the paragraph I the different types of standardization that we will consider, take intimately into account the structure constraints of the comparison between partitions.

3. LIMIT FORM OF $Q(\pi,\times)$. NORMALIZATION BY STANDARD DEVIATION

The solution that we will present [Lerman(1987c)] below concerns the more general context of comparing two symmetric (resp. antisymmetric) codings of OxO or weightings on OxO. If $\{\phi(x,y)/(x,y)\in XxX\}$ and $\{\psi(x,y)/(x,y)\in XxX\}$ -where $X=\{1,2,...,x,...,n\}$ labels O- denote the two codings, we suppose to have :

$$[\forall (x,y)\in XxX]\ [\phi(x,y)=\phi(y,x) \text{ and } \psi(x,y)=\psi(y,x)]$$

or

$[\forall (x,y)\in XxX]\ [\phi(x,y)=-\phi(y,x)$ and $\psi(x,y)=-\psi(y,x)]$.

On the other hand,
$(\forall x\in X)\ [\phi(x,x)=\psi(x,x)=1]$ in case of symmetric coding
or
$(\forall x\in X)\ [\phi(x,x)=\psi(x,x)=0]$ in case of antisymmetric coding.

The raw index does not take any account of the diagonal terms. It can be put in the following form :

$$s(\phi,\psi)=\sum\{\phi(x,y)\psi(x,y)/(x,y)\in X^{[2]}\}. \quad (1)$$

where $X^{[2]}=\{(x,y)/1\leq x\neq y\leq n\}$.

The random raw index can be put in the following form
$$S=s(\phi^*,\psi^*)=\sum\{\phi[\sigma(x),\sigma(y)]\psi[\tau(x),\tau(y)]/(x,y)\in X^{[2]}\} \quad (2)$$
where σ and τ are two independent random permutations taken in the set G_n of the n! permutations on X, provided by an uniform probability measure.

The comparison between two partitions π and $\times$ corresponds to that of two symmetric codings with 0-1 values and with the common value 1 on the diagonal of XxX. We have :

$$\mathrm{var}(S)=\frac{1}{n^{[2]}}\Big[\sum_{[x,y]}\phi(x,y)\Big]\Big[\sum_{[x,y]}\psi(x,y)\Big], \quad (3)$$

where $n^{[2]}=n(n-1)$ and $[x,y]\in X^{[2]}$.

In [Lerman(1987c)] we begin by comparing the variance expression obtained by N. Mantel (1967) with this one that we have set up completely independently [Lerman(1976)]. In fact we ignored the Mantel paper and the principle of our combinatorial calculation is rather different. It had been considered to make clear treatment after Lecalvé consideration (1976) highly inspired by an old paper of H.E. Daniels (1944). In our earliest reference we develop two new and elegant expressions of var(S). One of them enables us to determine very precisely the tendency of var(S) when n tends to infinity and when regular asymptotic conditions hold. Therefore, we may determine the limit form of $Q(\phi,\psi)$ that we present below. Let us introduce the following absolute moments

$$p_1=\frac{1}{n^2}\sum\{\phi(x,y)/(x,y)\in XxX\},$$

$$p_2=\frac{1}{n^2}\sum\{\phi^2(x,y)/(x,y)\in XxX\}, \quad (4)$$

$$t=\frac{1}{n^3}\sum\{\phi(x,y)\phi(x,z)/(x,y,z)\in XxXxX\},$$

on the other hand, respectively, q_1, q_2 and u associated to ψ in the same way as p_1, p_2 and t are associated to ϕ. Let us also consider

$$w=\frac{1}{n^2}\sum\{\phi(x,y)\psi(x,y)/(x,y)\in XxX\} \quad (5).$$

In the following theorem, we assume n tending to infinity, p_1,p_o,t (resp.q_1,q_o,u) and w tending to finite limits.

Theorem 1. The limit form of the standardized coefficient $Q(\phi,\psi)$ is :

$$Q(\phi,\psi)=\frac{\frac{\sqrt{n}}{2}(w-p_1q_1)}{\sqrt{(t-p_1^2)(u-q_1^2)+\frac{1}{2n}[(p_2-p_1^2)-2(t-p_1^2)][(q-q_1^2)-2(u-q_1^2)]}} \quad (5)$$

In general $(t-p_1^2)(u-q_1^2)$ is positive and different from zéro. In this general case, the expression of $Q(\phi, \psi)$ becomes for n "rather large" :

$$Q(\phi,\psi) \simeq \frac{\frac{\sqrt{n}}{2}(w-p_1q_1)}{\sqrt{(t-p_1^2)(u-q_1^2)}} \qquad (6)$$

It is of importance to notice that -as in the case of the construction of K.Pearson's coefficient- $Q(\phi,\psi)$ is to the multiplicative factor $\sqrt{n}$, a pure coefficient whose limit being independent of n.

Let us give now the limit form of the expression of $Q(\pi,\times)$ in case of comparison of two partition relations. In this case ϕ and ψ are symmetric with 0-1 values : $\phi(x,y)=1$ [resp. $\psi(x,y)=1$] if x and y are joined by the partition π(resp.$\times$) and $\phi(x,y)=0$ [resp. $\psi(x,y)=0$] if x and y are disjoined by the partition π(resp.$\times$). Reconsidering the notations introduced in paragraph II, let us set :

$$\pi_i = \frac{a_i}{n}, \quad x_j = \frac{b_j}{n} \text{ et } \gamma_{ij} = \frac{c_{ij}}{n} \text{ for every } (i,j),\ 1 \leqslant i \leqslant I,\ 1 \leqslant j \leqslant J.$$

Then we have

$$p_1 = p_2 = p = \sum_i \pi_i^2,\ t = \sum_i \pi_i^3,$$

$$q_1 = q_2 = q = \sum_j x_j^2,\ u = \sum_j x_j^3 \qquad (7)$$

$$\text{et } w = \sum_{(i,j)} \gamma_{ij}^2.$$

Theorem 2. (corollary of the theorem 1). The limit form of the standardized coefficient $Q(\pi,\times)$ is :

$$Q(\pi,\times) \simeq \frac{\frac{\sqrt{n}}{2}(w-pq)}{\sqrt{(t-p^2)(u-q^2)+\frac{1}{2n}\left[(p-p^2)-2(t-p^2)\right]\left[(q-q^2)-2(u-q^2)\right]}}. \qquad (8)$$

$(t-p^2)$ [resp.$(u-q^2)$] is in general strictly positive and exceptionally null ; the nullity occurs if the π_i (resp.x_j) are equal [i.e. $\pi_i=1/I$ for every i=1,2,...,I (resp. $x_j=1/J$ for every j=1,2,...,J)]. In this last particular case the magnitude order of the denominator changes and we have the impression of breaking in the behavior of the index $Q(\pi,\times)$. But we must keep in mind that the magnitude order of the numerator changes equally : the raw index $s(\pi,\times)$ falls in an abrupt way if one of the two partitions has their classes with the same cardinal, with respect to the situation where each of the two partitions has its cardinal classes very different mutually.

Now, let us come back to the most general case concerning the coefficient $Q(\phi,\psi)$. We have considered the possibility to define a coefficient with the following form :

$$R(\phi,\psi) = \frac{Q(\phi,\psi)}{\sqrt{Q(\phi,\phi)Q(\psi,\psi)}} \qquad (9)$$

If n is not large enough, in order to keep some influence to the second term under the square root sign ($\sqrt{\ }$) of (8) ; then we have clearly got a new coefficient. If not, the index that we obtain by this way corresponds exactly to the correlation coefficient between the two weightings ϕ and ψ, formally equivalent -but at the level of OxO- to K. Pearson's coefficient.

4. NORMALIZATION BY THE MAXIMUM

Our purpose is the elaboration of a coefficient consistent with $\tau(\alpha,\beta)$ [cf. expression (6) §2]. Relative to the comparison of two relational variables ϕ and ψ - corresponding to two codings of $O^{[2]}$ - the problem of the maximization of $\sum\{\phi(x,y)\psi(x,y)/(x,y)\in X^{[2]}\}$ on the set $\{\sum\{[\sigma(x),\sigma(y)]\psi(x,y)/(x,y)\in X^{[2]}\}/\sigma\in G_n\}$ -where G_n is the whole set of n! permutations on X- is recognized as a very difficult problem, whatever the types of the structures to be compared [Hubert(1983), Hubert & Arabie(1985), Marcotorchino (1984)].

We have been able to propose an exact solution to this problem in two situations. The first -which is of concern here- is of comparing two partitions π and $\times$ [Lerman and Peter(1986)]. The second is of comparing two total preorders [Lerman(1987a)]. The first situation is the more difficult to solve. We are going to give a brief look on the solution, the interested reader will refer to the detailed research report mentioned above.

According to the previous notations (cf.§1), it is in question to resolve in integer numbers :

$$\text{Max } \sum\{c_{ij}^2 / 1\leqslant i\leqslant I,\ 1\leqslant j\leqslant J\} \qquad (1)$$

under the following constraints

$$\begin{aligned} &\sum_{1\leqslant j\leqslant J} c_{ij} = a_i \text{ for all } i,\ 1\leqslant i\leqslant I \\ &\sum_{1\leqslant i\leqslant I} c_{ij} = b_j \text{ for all } j,\ 1\leqslant j\leqslant J \end{aligned} \qquad (2)$$

The first very important idea is to replace the notion of 'mathematical formulae' -tried till now- by the one, much more general of 'recursive algorithm'. The second idea -linked to the first- consists of working at the level of the contingency table (I rows and J columns). We start by filling up the margins $\{a_i/1\leqslant i\leqslant I\}$ and $\{b_j/1\leqslant j\leqslant J\}$ that we have to distribute in the inside of the contingency table, in a compatible way and in order to maximize $\sum\{c_{ij}^2/1\leqslant i\leqslant I,\ 1\leqslant j\leqslant J\}$. Moreover, we will define by this way an optimal configuration of the contingency table.

We may denote -without ambiguity- by π(resp.$\times$) the indicator function of $R(\pi)$ [resp. $R(\times)$] in $P = O^{\{2\}}$ (cf. notations of the paragraph I). We establish (cf. above mentioned reference) that among the classical bounds majoring (1) by the means of mathematical expression, symmetric with respect π and $\times$, the best one (i.e. the lowest) is obtained by application of the Shwartz inequality conceived in a logical framework:

$$\sum_{p\in P} [\pi(p)-\mu][\times(x)-\nu] \leqslant \sqrt{\{(\sum_{p\in P} [\pi(p)-\mu]^2)(\sum_{p\in P} [\times(p)-\nu]^2)\}} \qquad (3)$$

$$\text{where } \mu = \sum_i a_i(a_i-1)/n(n-1) \text{ and } \nu = \sum_j b_j(b_j-1)/n(n-1). \qquad (4)$$

Nevertheless, we show this 'analytical' bound for $\sum\{c_{ij}^2/1\leqslant i\leqslant I,\ 1\leqslant j\leqslant J\}$ to be too large with respect to the non symmetric one defined by $\min(\sum_i a_i^2, \sum_j b_j^2)$. This last becomes the exact bound if one of the two partitions $\{a_i/1\leqslant i\leqslant I\}$, $\{b_j/1\leqslant j\leqslant J\}$ is finer than the other one.

The recursive solution is based on the fact that facing to an optimal configuration, if we delete a row (resp. column) and we consequently fit the margins, we also obtain an optimal configuration.

Relative to the contingency table, empty in its inside, but having full margins, we will be led -at each step- to fit up the content of a row margin (resp. or a column

margin), that we denote by α_i (resp.β_j), in a column j (resp. row i) such that $\beta_j \geq \alpha_i$ (resp.$\alpha_i \geq \beta_j$) ; we will say that we 'resolve' the couple (i,j). Such 'resolution' -considered in the framework of the optimal configuration- decreases the problem size by decreasing by one unity or even by two (if the two components of the resolved couple are identical) the cardinal defined by (number of rows + number of columns).

We begin by showing that the greatest integer $c_{i_0 j_0}$ of the optimal configuration T_0 of the contingency table T, corresponds necessarily to the resolution of the couple (i_0,j_0). On the other hand, we show that if the same integer is in row margin and in column margin ($a_{i_1}=b_{j_1}$), the optimal configuration T_0 includes necessarily the resolution of (i_1,j_1).

A deep experimental analysis led us to introduce on the set of the couples $\{(a_i,b_j)/1 \leq i \leq I,\ 1 \leq j \leq J\}$, a relation of partial preorder, resulting from the intersection of two total preorders ω_d and ω_s, where ω_d is consistent with a decreasing 'difference' $|a_i-b_i|$ and where ω_s is consistent with an increasing sum (a_i+b_j). More precisely,

$\{\forall [(i,j),(i',j')]\},\ (a_{i'},b_{j'}) \leq (a_i,b_j)$ (for ω_d) $\Leftrightarrow$ $|a_i-b_j| \leq |a_{i'}-b_{j'}|$ (5)

and

$\{\forall [(i,j),(i',j')]\},\ (a_{i'},b_{j'}) \leq (a_i,b_j)$ (for ω_s) $\Leftrightarrow$ $(a_i+b_j) \geq (a_{i'}+b_{j'})$. (6)

The algorithm that we propose is based on the following simple property : "An optimal configuration of the table T can be obtained by starting with the resolution of an extremal couple (a_i,b_j), with respect to $\omega=\omega_d \cap \omega_s$".

The major result that we have been able to establish [Lerman & Peter(1986)] consists of the complete demonstration of this property, in case where it exists only one extremal couple. On the other hand we have realized in different mathematical situations where it exists more than one extremal couple, that the optimal solution goes necessarily through resolution of an extremal couple. A counter example to this property would be highly unliked ; it would express that none of the resolved couples at the level of an optimal configuration, correspond to an extremal couple. But, on the one hand, the resolution of an extremal couple corresponds to empty a margin in order to fill up for the best the inside of the table. On the other hand, we know that -at each step- the greatest entry of an optimal configuration (reduced step by step) corresponds to a resolution of a couple.

The process of recursive research involves the determination -after each resolution- of the set of the extremal couples. The following important property bounds extensively the stacking :

If (a_{i_Q},b_{j_Q}) is the extremal couple for which $|a_i-b_j|$ is minimal, the unloading of a_{i_0} in j_0 (if $a_{i_0} \leq b_{j_0}$) [resp. of b_{j_0} in i_0 (if $a_{i_0} \geq b_{j_0}$)] preserves the extremal character of any other couple (a_{i_1},b_{j_1}) with $i_1 \neq i_0$ $j_1 \neq j_0$.

Thus to obtain the coefficient $\tau(\pi,\times)$ according to the formulae (6) of paragraph 2, we do not have to use a mathematical formulae which -in any case- would give a poor bound. But we may use the algorithm that we have presented above and which is detailed in [Lerman and Peter(1986)].

5. STATISTICAL ASPECTS

We have emphasized the non relevance of testing statistical independence hypothesis between qualitative variables in data analysis. What is more in question is the mutual organization of a variable family according to their respective mutual relations. Our hierarchical classification method based on the likelihood of the maximal association (or link) [Lerman(1970a),(1981)] provides a solution to this problem. In the framework of this paper let us consider the case where the data is a family $\{\pi_l / 1 \leq l \leq L\}$ of partitions. Suppose $\{Q(\pi_l,\pi_m)/1 \leq l < m \leq L\}$ or $\{R(\pi_l,\pi_m)/1 \leq l < m \leq L\}$ the association coefficient matrix established according to the diagram and expressions (1) (§2), or

respectively to (9) (§3). Letting $\{S(\pi_l,\pi_m)/1\leq l<m\leq L\}$ denote one of the two preceding tables. The passage from this last to the table $\{P(\pi_l,\pi_m)/1\leq l\leq m\leq L\}$, where $P(\pi_l,\pi_m)$ concerns probability scale defined by the likelihood of the link, is done by means of the formulae :

$$P(\pi_l,\pi_m)= \Phi\left[\frac{S(\pi_l,\pi_m)-moy_e(S)}{\sqrt{var_e(S)}}\right] \quad , 1\leq l<m\leq L, \qquad (1)$$

where Φ is the cumulative function of the normal $[N(0,1)]$ distribution and where $moy_e(S)$ and $var_e(S)$ are respectively the mean and variance of the table of values $\{S(\pi_l,\pi_m)/1\leq l<m\leq L\}$.

Although the reference (1) to the normal distribution is always algorithmically possible, a better justification requires normal tendency of the random index $Q(\pi^*,\times^*)$. However, P.W. Mielke (1979) showed non normal tendency in case of partitions having classes with common cardinal. But partitions having equal size classes correspond to a pure human construction, they cannot be found in natural data.

REFERENCES

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

[2] Fowlkes, E.B. and Mallows C.L. (1983). 'A method for comparing two hierarchical clusterings', J.A.S.A. 78 553-569.

[3] Goodman L.A. and Kruskal W.H. (1954), 'Measures of association for cross classifications', J.A.S.A. 49, 732-764.

[4] Hamann V. (1961), 'Merkmalbestand und Verwandtschaftsbeziehungen der Farinosae. Ein Beitragzum System der Monokotyledonen', Willdenowia, 2, 639-768.

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

[6] Hubert L.J. and Arabie Ph. (1985), 'Comparing partitions', Journal of Classification, 2,2-3, 193-218.

[7] Jaccard P. (1908), 'Nouvelles recherches sur la distribution florale', Bull. Soc. Vaud. Sci Nat., t.44, 223-270.

[8] Kendall M.G. (1970), 'Rank correlation methods', Charles Griffin, fourth edition (first edition in 1948).

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

[10] Lerman I.C. (1970a), 'Sur l'analyse des données préalable à une classification automatique. Proposition d'une nouvelle mesure de similarité', Rev. Math. & Sc. Hum. 8è année n°32.

[11] Lerman I.C. (1970b), 'Les bases de la classification automatique', Gauthier-Villars, Paris.

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

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

[14] Lerman I.C. (1981), 'Classification et analyse ordinale des données', Dunod, Paris.
[15] Lerman I.C. (1983a), 'Indices d'association partielle entre variables qualitatives nominales', R.A.I.R.O., série R.O., vol.17, n°3, 213-259.
[16] Lerman I.C. (1983b), 'Indices d'association partielle entre variables qualitatives ordinales', Publ. I.S.U.P., XXVIII, fasc 1,2, 7-46.
[17] 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. ISUP, XXIX, fasc. 3-4, 27-57.
[18] Lerman I.C. (1987a), 'Maximisation de l'association entre deux variables qualitatives ordinales', Publ. Int. n°341, IRISA, Rennes.
[19] Lerman I.C. (1987b), 'Comparing relational variables according to Likelihood of the links classification method', in Recent Developments in Clustering and Data Analysis, Japanese-French Scientific Seminar, 24-26 March 1987, Tokyo, to be published by Academic Press.
[20] Lerman I.C. (1987c), 'Analyse formelle de coefficients statistiques d'association entre variables relationnelles', Publ. Int. à paraître, IRISA, Rennes.
[21] Lerman I.C. et Peter Ph. (1986), 'Structure maximale pour la somme des carrés d'une contingence aux marges fixées. Une solution algorithmique programmée', Publ. Int. n°318, IRISA, Oct. 90 pages, Rennes.
[22] Mantel N. (1967), 'The detection of disease clustering and a generalized regression 'approach', Cancer Research, vol.27, n°2, 209-220.
[23] Marcotorchino F. (1984), 'Utilisation des comparaisons par paires en statistique des contingences (Partie I)', Etude F-069, Centre Scientifique IBM-France.
[24] Mielke P.W. (1979), 'On asymptotic non normality of null distributions of MRPP Statistics', Communications in Statistics, Thoery and Methods, A8(15), 1541-1550.
[25] Ochiai A. (1957), 'Zoogeographic studies on the soleoid fishes found in Japan and its neighbowring regions', Bull. Jap. Soc. Sci. Fish, T22, 526-530.
[26] Pearson K. (1926), 'On the coefficient of racial likeness', Biometrika t.18, 105-117.
[27] Rand W.M. (1971), 'Objective criteria for the evaluation of clustering methods', J.A.S.A. 66, 846-850.
[28] Sokal R.R. and Michener C.D. (1958), 'A statistical method for evaluating systematic relationships', Univ Kansas Sci. Bull. 38, 1409-1438.
[29] Yule G.U. (1911), 'An introduction of the theory of statistics', Charles Griffin and Co. Ltd, London.
[30] Yule G.U. (1912), 'On the methods of measuring the association between two attributes', J. Roy. Statist. Soc. 75, 579-652.