Justification et validité statistique d'une echelle [0,1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées

Israël-César Lerman

▶ **To cite this version:**

Israël-César Lerman. Justification et validité statistique d'une echelle [0,1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées. [Rapport de recherche] RR-0278, INRIA. 1984. inria-00076280

HAL Id: inria-00076280
https://inria.hal.science/inria-00076280

Submitted on 24 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

IRIA

CENTRE DE RENNES
IRISA

Rapports de Recherche

N° 278

JUSTIFICATION ET VALIDITÉ STATISTIQUE D'UNE ÉCHELLE [0,1] DE FRÉQUENCE MATHÉMATIQUE POUR UNE STRUCTURE DE PROXIMITÉ SUR UN ENSEMBLE DE VARIABLES OBSERVÉES

Israël César LERMAN

Mars 1984

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél. (3) 954 90 20

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTÈMES ALÉATOIRES

Campus Universitaire de Beaulieu
Avenue du Général Leclerc
35042 - RENNES CÉDEX
FRANCE
Tél. : (99) 36.20.00
Télex : UNIRISA 95 0473 F

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

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

Publication Interne n°221
Janvier 1984
47 pages

Résumé : Ce travail étudie et précise la situation de notre approche dans l'élaboration d'un coefficient d'association entre variables qualitatives par rapport à un cadre d'inférence statistique plus classique introduit par L.A. Goodman et W.H. Kruskal, où l'ensemble E des individus (de cardinal n) est regardé comme l'observation d'un échantillon aléatoire pris dans une population $\mathcal{P}$ de très grande taille.
Après une réflexion sur la comparaison entre l'optique de l'analyse des données et celle des tests non paramétriques d'hypothèses, nous montrons comment nos "hypothèses d'absence de liaison" ("h.a.l.") -qui ont un caractère intrinsèque : au niveau de E et sans aucune référence à $\mathcal{P}$ - sont essentielles pour construire l'expression formelle des indices d'association. On peut alors déterminer le comportement limite (pour n -> ∞) de cette expression formelle, ce qui permet de préciser ces indices -sans arbitraire ni biais- au niveau de la population parente $\mathcal{P}$.
C'est alors qu'on peut étudier la distribution dans l'optique inférentielle des indices aléatoires d'association réalisés sur E. Ce que nous faisons en détail dans le cas de la comparaison d'attributs descriptifs et qui nous permet d'avoir un avis sur la précision calcul pertinente dans une analyse des données.
Cette vue inférentielle nous permet aussi de justifier la pratique de la réduction globale des similarités, avant la référence à une échelle $[0,1]$ de fréquence mathématique dite de "vraisemblance des liaisons observées".
Ce faisant, nous sommes conduits à aborder un problème fondamental de la statistique non paramétrique multivariée, qui est celui de la distribution de la table des indices aléatoires d'association entre variables.

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
(L.A. 227)
UNIVERSITÉ DE RENNES 1 I.N.S.A. DE RENNES

INSTITUT NATIONAL DE RECHERCHE
EN INFORMATIQUE ET EN AUTOMATIQUE
(LABORATOIRE DE RENNES)

Summary : In the terms of "hypothesis of non link" (h.n.l.), our approach in establishing the association coefficients between qualitative variables formalizes, generalizes and uses directly free distribution of permutational type provided by non parametric statistical inference. The first aspect of this work is to analyse the position of this approach with respect to the statistical inference hypothesis of another type studied by L.A. Goodman and W.H. Kruskal. In their paper, these authors consider the set E of individuals (card(E)=n) as an observation of random sample from large population $\mathcal{P}$.
After looking at the difference between approaches of data analysis and non parametric statistical hypothesis, we show why the "h.n.l." -which is defined at the level of E without referring to $\mathcal{P}$ - is essential to determine the formal expression of the association indices. Formal expression may be transposed at the level of $\mathcal{P}$ and then, it becomes relevant to study the distribution of the random indices observed on E. This type of study allows us to indicate the necessary limitation of the significant precision in the computing of the indices on E.
On the other hand, the statistical inference point of view -where E is an observed random sample of $\mathcal{P}$ - justifies the validity of the global normalization of the indices of association between descriptive variables. This standardization precedes the reference to a $[0,1]$ scale of the "likelihood of the observed relations". Hence, we are led to consider the fundamental problem of free distribution of the random table of proximity indices between variables.

PAPIER RÉCUPÉRÉ ET RECYCLÉ

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

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

I. INTRODUCTION

Notre but est ici -conformément à une question qui nous a souvent été posée- de chercher à situer le principe de notre approche dans l'élaboration d'un indice d'association entre variables qualitatives, par rapport à un cadre d'inférence statistique plus classique. Ce cadre est celui où l'ensemble E des objets ou individus, au niveau duquel on travaille, se trouve regardé comme résultant d'un échantillonnage aléatoire dans une population $\mathcal{P}$ de très grande taille, n et N désigneront les cardinaux respectifs de E et $\mathcal{P}$.

Notre augmentation se réfèrera aux deux optiques classiques ; d'une part et principalement à celle des tests d'hypothèses d'absence de liaison entre variables, d'autre part et de façon d'ailleurs nécessairement liée, à l'aspect estimation. Ce dernier aspect a été développé extensivement dans les travaux de L. Goodman et W. Kruskal [GOODMAN et KRUSKAL(1963),(1972)], relativement aux indices que ces auteurs proposent. Toutefois, les bases formelle et statistique de l'expression de ces derniers indices -et on s'en rend compte lorsqu'il s'agit de les définir au niveau de la population parente dont provient l'échantillon sur lequel les variables ont été effectivement observées- ne sont pas clairement établies. Nous y reviendrons.

Pour demeurer sur le plan des principes en ne nous encombrant pas d'aspects techniques, nous allons considérer la situation la plus simple de la comparaison d'attributs descriptifs (i.e. variables logiques 0-1, d'absence-présence). Nous verrons que le problème est différent selon qu'il s'agit d'évaluer l'association entre deux attributs, de comparer les valeurs des associations pour deux paires d'attributs ou surtout, pour toutes les paires d'un même ensemble $\mathcal{A}$ d'attributs de description d'un champ fixé. C'est cette dernière structure de proximité qui nous intéresse au premier chef pour la construction ascendante hiérarchique d'un arbre de classification sur un ensemble de variables descriptives ; laquelle permet de dégager les principales tendances comportementales de la population étudiée à travers un échantillon.

II. POPULATION, ECHANTILLON, ANALYSE DES DONNEES ET TESTS D'INDEPENDANCE

Le titre de ce paragraphe correspond à un programme par trop ambitieux ! Nous avons quant à nous besoin d'exprimer dans ce cadre quelques remarques intuitives très simples.

Nous pensons que la philosophie de l'"analyse des données" est en quelque sorte opposée à celle des "tests de l'hypothèse d'indépendance ou d'absence de liaisons". En effet, prenons par exemple le problème de la recherche de liaisons sur $\mathcal{P}$ entre éléments d'un ensemble V de variables. La deuxième optique (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 $\mathcal{P}$, 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. Cette évaluation va quand même devoir s'effectuer sur la base de l'échantillon E qu'on espère "représentatif" de $\mathcal{P}$. 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'association entre variables. Toutefois, pour des liaisons "faibles" et un effectif n de l'échantillon (n=card(E)) non "assez grand", 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. [BLANCARD(1976)].

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 vons considéré une suite croissante d'échantillons aléatoires dont la suite des tailles est la suite des multiples entiers de 1000 (1000, 2000, 3000, ...). On a pu constater que les classes hiérarchiques d'attributs qui étaient bien structurées (marquées par des "noeuds significatifs" dans notre méthode) et correspondaient à des liaisons nettes, se trouvaient préservées. D'autres classes correspondant à des profils 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 restaient 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), (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.

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 & WOLFOWITZ (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 $(x_1, x_2, \ldots, x_i, \ldots, x_n)$ de la variable v (resp. $(y_1, y_2, \ldots, y_i, \ldots, y_n)$ de la variable w) - où $\{1, 2, \ldots, i, \ldots, n\}$ code E - la suite aléatoire $(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(i)}, \ldots, x_{\sigma(n)})$ (resp. $(y_{\tau(1)}, y_{\tau(2)}, \ldots, y_{\tau(i)}, \ldots, y_{\tau(n)})$ où σ (resp. τ) est un élément aléatoire dans l'ensemble G_n, muni d'une probabilité uniforme des n! permutations sur $(1, 2, \ldots, i, \ldots, n)$. On situe ensuite l'indice $\mathcal{E}\{x_i\, y_i \;/\; 1 \leq i \leq n\}$ (que nous appelons "brut") par rapport à la distribution commune de l'une ou de l'autre des deux variables aléatoires (v.a.) : $\mathcal{E}\{x_i\, y_{\tau(i)} \;/\; 1 \leq i \leq n\}$ et $\mathcal{E}\{x_{\sigma(i)}\, y_i \;/\; 1 \leq i \leq n\}$. L'association $(1, 2, \ldots, i, \ldots, n) \longrightarrow (\sigma(1), \sigma(2), \ldots, \sigma(i), \ldots, \sigma(n))$ (resp. $(\tau(1), \tau(2), \ldots, \tau(i), \ldots, \tau(n))$ est ce que nous désignons par "hypothèse d'absence de liaison" (h.a.l.).

On verra au paragraphe III.1. ci-dessous ce que devient l'expression de cette "h.a.l." lorsqu'il s'agit de comparer deux variables "attributs". Nous avons d'autre part pu voir que l'extension nécessaire [LERMAN (1976), (1981)] de cette forme de l'hypothèse d'absence de liaison à la comparaison de deux variables qualitatives nominales (resp. ordinales) consiste à remplacer la notion de permutation aléatoire par celle de partition (resp. préordre total) aléatoire. De façon précise, à la partition (resp. préordre total) observé sur E , on associera une partition (resp. préordre total) aléatoire dans l'ensemble, muni d'une probabilité uniforme, des partitions sur E de même type cardinal que celui de la partition donnée (resp. des préordres totaux sur E de même composition que celle observée).

Signalons au passage qu'en partant d'un indice brut conçu au niveau de E x E [LERMAN (1973), (1981)], on peut selon un principe qui peut se déduire de celui du test de permutation de Wald et Wolfowitz, construire un test d'absence de liaison qui ne fait nullement référence à la distribution jointe des deux variables qualitatives sur la population totale et parente

Cependant, nous utilisons quant à nous la distribution des v.a. associées aux indices bruts dans l'optique de l'analyse des données (organisation comparée des liaisons), optique que nous avons présentée comme contraire à la démarche des tests d'hypothèses.

Comme nous l'avons annoncé, nous allons pour mettre clairement en avant les idées, considérer la situation techniquement la plus simple de la comparaison d'attributs.

III. ASSOCIATION ENTRE DEUX ATTRIBUTS

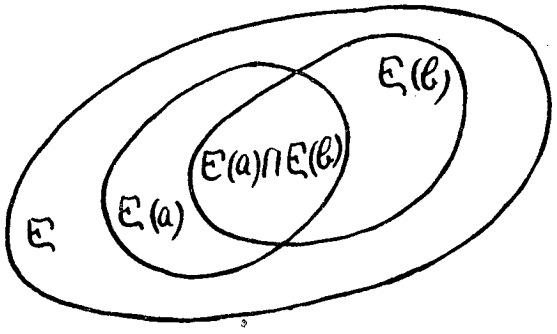
Rappelons que E est l'ensemble des individus - de cardinal n - défini par l'échantillon qu'on espère représentatif de la population $\mathcal{P}$ qui définit un ensemble de cardinal N .

Nous avons mis en évidence un schéma très général pour la construction d'un indice d'association entre deux variables statistiques mesurées sur E quelle que soit leur nature combinatoire LERMAN (1981), (1983) . Le point de départ de cette construction a d'ailleurs été pour nous la comparaison de deux attributs descriptifs a et b, comparaison que nous allons reprendre avec la forme la plus simple de l'hypothèse d'absence de liaison.

III.1. INDICE CENTRE REDUIT, INDICE DE LA VRAISEMBLANCE DU LIEN, AU NIVEAU DE E.

Dans ce sous paragraphe aucune référence n'est faite à la population environnante $\mathcal{P}$; tout se passe comme si notre seul univers est l'ensemble E .

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



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

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

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

$$Q(a,b)=\sqrt{(n-1)}\ \times\ \frac{\left[p(a\wedge b)p(\bar{a}\wedge\bar{b})-p(a\wedge\bar{b})p(\bar{a}\wedge b)\right]}{\sqrt{p(a)p(\bar{a})p(b)p(\bar{b})}} \qquad (1)$$

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

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

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

$$P(a,b)=Pr\{S(a)\leqslant s/N_1\}=Pr\{S(b)\leqslant s/N_1\} \qquad (2)$$

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

Dans un test non paramétrique d'hypothèse -défini au niveau de E (cf. § II ci-dessus)- on ne se sert de la probabilité $Pr\{S>s/N_1\}$ que pour rejeter ou non l'hypothèse d'absence de lien entre les deux attributs a et b. Alors que nous prétendons utiliser de façon beaucoup plus riche cette échelle de probabilité pour en faire véritablement une échelle de mesure de la liaison. Ceci est parfaitement conforme à l'optique offensive de l'analyse des données. mais n'est pas sans risque de glissement dans les notions et de précision dans les estimations calculées des ressemblances. L'objet de notre réflexion ici consiste précisément à montrer comment neutraliser ces deux risques et obtenir de la sorte une échelle très fine et très riche pour la comparaison deux à deux d'un ensemble de variables, ce qui conduit au critère très général de la "vraisemblance du lien" pour la construction ascendante hiérarchique d'un arbre des classifications sur l'ensemble des variables de description.

Revenons à l'expression (1) ci-dessus et mettons-la sous la forme

$$Q(a,b)=\sqrt{n-1}\ \ r(a,b) \qquad (3)$$

où

$$r(a,b)=\left[p(a\wedge b)p(\bar{a}\wedge\bar{b})-p(a\wedge\bar{b})p(\bar{a}\wedge b)\right]\Big/\sqrt{p(a)p(\bar{a})p(b)p(\bar{b})} \qquad (4)$$

qui est le coefficient de K. Pearson, peut être interprété comme le coefficient de corrélation entre les deux variables α et β où $\alpha(i)=1$ ou 0 (resp. $\beta(i)=1$ ou 0) selon que l'attribut a(resp. b) est présent ou non chez le i-ème individu, $1\leqslant i\leqslant n$. On a

$$-1\leqslant r(a,b)\leqslant 1 \qquad (5)$$

En ce qui concerne l'expression (2) ci-dessus, compte tenu de l'excellente approximation normale de la loi hypergéométrique -pourvu que $p(a)\wedge p(b)$ ne soit pas trop voisin de 0(resp. 1) et que n ne soit pas trop petit (ce qui est le cas général en analyse de données)- sa valeur calculée sera approchée par

$$\Phi\left[\sqrt{(n-1)}\ r(a,b)\right] \qquad (6)$$

où Φ est la fonction de répartition (f.r.) de la loi normale centrée réduite. La qualité de cette approximation n'aura pas à être discutée.

III.2. COMPARAISON DE DEUX ATTRIBUTS PAR RAPPORT A UN ECHANTILLON DE TAILLE CROISSANTE

III.2.1. Introduction ; les différentes expressions d'un indice d'association

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

On peut au niveau de $\mathcal{P}$ définir et d'ailleurs, de la même façon qu'au niveau de E, l'indice correspondant à r(a,b) :

$$\rho(a,b)=\left[\pi(a\wedge b)\pi(\bar{a}\wedge\bar{b})-\pi(a\wedge\bar{b})\pi(\bar{a}\wedge b)\right]\Big/\sqrt{\pi(a)\pi(\bar{a})\pi(b)\pi(\bar{b})} \qquad (7)$$

S'il s'agit de donner une "mesure" de l'indice d'association entre les deux attributs a et b au niveau de l'ensemble E -en "oubliant" $\mathcal{P}$ - on peut proposer l'indice r(a,b) ou celui de la "vraisemblance du lien" P(a,b) calculé au moyen de la formule (6) ci-dessus avec l'intéressante interprétation qui accompagne ce dernier indice qui reste une fonction croissante du premier.

Cependant, s'il s'agit de comparer les deux attributs a et b sur la base de l'échantillon aléatoire E, mais relativement à une optique "verticale ascendante" où la taille n augmente, on ne peut plus proposer l'indice P(a,b).

En effet, pour cette situation évolutive où n croît, les proportions calculées $p_n(a\wedge b)$, $p_n(a\wedge\bar{b})$, $p_n(\bar{a}\wedge b)$ et $p_n(\bar{a}\wedge\bar{b})$ -où n a été placé en indice pour rappeler la taille de l'échantillon- deviennent des estimations de plus en plus précises de, respectivement, $\pi(a\wedge b),\pi(a\wedge\bar{b}),\pi(\bar{a}\wedge b)$ et $\pi(\bar{a}\wedge\bar{b})$. De sorte que $r_n(a,b)$ devient une estimation de plus en plus précise de $\rho(a,b)$ (cf. formule (7)) et si $\rho(a,b)$ est strictement positif (resp. négatif)

$$P_n(a,b)=\Phi\left[\sqrt{(n-1)}\ r_n(a,b)\right] \qquad (8)$$

peut être rendu -pour n assez grand- aussi voisin de 1(resp.0) qu'on le veut! C'est que l'échelle de probabilité -définie ici par la loi normale- est dans

ce contexte juste conçue pour la mise en évidence d'un lien, indépendamment de sa mesure effective.

Cette mesure effective ne peut être définie au niveau de la population qu'à partir des proportions $\pi(a\wedge b)$,$\pi(a\wedge\bar{b})$,$\pi(\bar{a}\wedge b)$ et $\pi(\bar{a}\wedge\bar{b})$. Toutefois, cette mesure ne peut être posée a priori comme cela est fait, relativement à l'association entre variables qualitatives dans [GOODMAN & KRUSKAL(1958)], elle doit résulter d'une étude statistique non paramétrique de même type que celle du paragraphe III.1., mais transposée au niveau de $\mathcal{P}$. Dans ces conditions, il est naturel et justifié de retenir l'indice $\rho(a,b)$.

Cependant, l'expression (7) ci-dessus ne correspond pas au seul indice obéissant au principe qu'on vient d'exprimer. Nous avons en effet pu mettre en évidence [LERMAN(1981)] trois formes fondamentales de l'hypothèse d'absence de liaison pour la comparaison de deux attributs descriptifs a et b au niveau de l'ensemble E où ils sont observés (construction de même type que celle du paragraphe III.1.). Ces trois formes se distinguent de la manière plus ou moins diffuse dont elles respectent les cardinaux des sous ensembles E(a) et E(b) (cf. § III.1.).

L'indice présenté ci-dessus correspond comme nous l'avons déjà précisé à la forme appelée N_1 de l'h.a.l., c'est pour cette raison que nous le désignerons par ${}_1r_n(a,b)$, ou plus simplement par ${}_1r(a,b)$. Ce même indice défini au niveau de la population $\mathcal{P}$ sera naturellement noté ${}_1\rho(a,b)$.

Pour la forme N_2 de l'h.a.l., on associe au couple (E(a),E(b)), un couple (X,Y) de parties aléatoires indépendantes de E, où le choix aléatoire de Y se fait selon un principe analogue à celui de X. Pour ce dernier choix, nous munissons l'ensemble $\mathcal{P}(E)$ des parties de E d'une mesure de probabilité plus diffuse que dans le cas de l'h.a.l. N_1 ; alors que pour le modèle aléatoire 1, la mesure de probabilité était concentrée sur le seul niveau du simplexe $\mathcal{P}(E)$ associé à $\mathcal{P}_{n(a)}(E)$, elle sera ici répartie sur les différents niveaux. Ainsi, le modèle aléatoire 2 comporte deux pas : le premier consiste dans le choix d'un niveau et le second, dans le choix d'un élément de ce niveau. Pour le choix du niveau, on introduit la v.a. K indice d'un même niveau et cardinal commun de toutes les parties de ce niveau. K est considérée comme une v.a. binomiale de paramètres (n,p(a)) où p(a)=n(a)/n. Pour le choix aléatoire d'un élément de même niveau k, la probabilité binomiale affectée à ce niveau est uniformément répartie sur l'ensemble des $\binom{n}{k}$ sommets de ce niveau (dont chacun représente une partie de cardinal k).

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

$$ {}_2Q(a,b)=\sqrt{n}\ {}_2r(a,b) \qquad (9)$$

où

$$ {}_2r(a,b)=\left[p(a\wedge b)p(\bar{a}\wedge\bar{b})-p(a\wedge\bar{b})p(\bar{a}\wedge b)\right]\Big/\sqrt{p(a)p(b)\left[1-p_{(a)}p_{(b)}\right]} \qquad (10)$$

Le dénominateur du rapport qui définit ce dernier indice étant plus grand que celui de l'indice ${}_1r(a,b)$ (cf. formule (4) ci-dessus), on a également

$$ -1\leqslant {}_2r(a,b)\leqslant 1. \qquad (11)$$

Dans l'h.a.l. N_2, pour n assez grand, on a l'excellente approximation normale de la loi discrète de ${}_2Q(x,y)$. De sorte que l'indice de la vraisemblance du lien calculé dans le cadre de E et relativement à N_2 (cf.§ III.1.), peut se mettre sous la forme

$$\Phi\left[\sqrt{n}\ {}_2r(a,b)\right] \qquad (12)$$

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

A l'indice ${}_2r(a,b)$ dégagé au niveau de E, correspond selon une induction ci-dessus argumentée l'indice défini au niveau de la population :

$${}_2\rho(a,b)=\left[\pi(a\wedge b)\pi(\bar{a}\wedge\bar{b})-\pi(a\wedge\bar{b})\pi(\bar{a}\wedge b)\right]\Big/\sqrt{\pi(a)\pi(b)\left[1-\pi(a)\pi(b)\right]} \qquad (13)$$

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

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

$${}_3Q(a,b)=\sqrt{n}\ {}_3r(a,b), \qquad (14)$$

où on a

$${}_3r(a,b)=\left[p(a\wedge b)p(\bar{a}\wedge\bar{b})-p(a\wedge\bar{b})p(\bar{a}\wedge b)\right]\Big/\sqrt{p(a)p(b)} \qquad (15)$$

Le dénominateur de ce rapport étant plus grand que celui du rapport (10) définissant ${}_2r(a,b)$, on a a fortiori :

$$-1\leqslant{}_3r(a,b)\leqslant 1 \qquad (16)$$

Ici encore, dans l'h.a.l. N_3, pour n assez grand, on a l'excellente approximation normale de la loi de ${}_3Q(x,y)$. De sorte que l'indice de la vraisemblance du lien, calculé au niveau de E et relatif à N_3, peut se calculer très approximativement par

$$\Phi\left[\sqrt{n}\ {}_3r(a,b)\right] \qquad (17)$$

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

Comme précédemment, à l'indice ${}_3r(a,b)$ dégagé au niveau de E, correspond par induction l'indice défini au niveau de la population :

$${}_3\rho(a,b)=\left[\pi(a\wedge b)\pi(\bar{a}\wedge\bar{b})-\pi(a\wedge\bar{b})\pi(\bar{a}\wedge b)\right]/\sqrt{\pi(a)\pi(b)} \qquad (18)$$

On constate que les numérateurs des trois indices ${}_1r(a,b)$, ${}_2r(a,b)$ et ${}_3r(a,b)$ (resp. ${}_1\rho(a,b)$, ${}_2\rho(a,b)$ et ${}_3\rho(a,b)$) sont identiques. Si ce numérateur est positif, on a

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

$$(\text{resp.}\ {}_3\rho(a,b)<{}_2\rho(a,b)<{}_1\rho(a,b)) \qquad (19)$$

Le modèle Binomial fournit donc un indice "intermédiaire" entre les indices Poissonnien et Hypergéométrique. Si pour l'analyse statistique, le modèle Binomial est le plus souple à manipuler, par contre pour l'analyse des données, le choix se présente le plus clairement entre les indices ${}_3r(a,b)$ et ${}_1r(a,b)$ (resp. ${}_3\rho(a,b)$ et ${}_1\rho(a,b)$).

L'indice ${}_1r(a,b)$ (resp. ${}_1\rho(a,b)$) est parfaitement symétrique : si a et b sont deux attributs rares pour lesquels le numérateur commun des indices est positif, on a

$$ {}_1r(a,b)={}_1r(\bar{a},\bar{b}) $$
$$ (\text{resp.}\ {}_1\rho(a,b)={}_1\rho(\bar{a},\bar{b})), \qquad (20) $$

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

$$ {}_3r(a,b)>{}_3r(\bar{a},\bar{b}) $$
$$ (\text{resp.}\ {}_3\rho(a,b)>{}_3\rho(\bar{a},\bar{b})) \qquad (21) $$

Le mathématicien épris de symétrie aura tendance à préférer l'indice ${}_1r$ (resp. ${}_1\rho$). Toutefois, en analyse des données où on travaille avec un ensemble $\mathcal{A}$ d'attributs "orientés" (i.e. $a\in\mathcal{A} \iff a\notin\mathcal{A}$), toutes choses "égales par ailleurs", il importe que l'association entre attributs rares soit plus ponctuée que celle entre attributs fréquents. C'est précisément ce que réalise le modèle N_3 de l'h.a.l.

III.2.2. Intervalles de confiance pour ${}_1\rho(a,b)$ et pour ${}_3\rho(a,b)$; conséquence sur la précision de calcul permise en analyse des données.

III.2.2.1. Modèle aléatoire d'échantillonnage de E et notations

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

$$ W=\left[\begin{array}{c|c|c|c}
\pi(a\wedge b)[1-\pi(a\wedge b)] & -\pi(a\wedge b)\pi(a\wedge\bar{b}) & -\pi(a\wedge\bar{b})\pi(\bar{a}\wedge b) & -\pi(a\wedge\bar{b})\pi(\bar{a}\wedge\bar{b}) \\
-\pi(a\wedge\bar{b})\pi(a\wedge b) & -\pi(a\wedge\bar{b})[1-\pi(a\wedge\bar{b})] & -\pi(a\wedge\bar{b})\pi(\bar{a}\wedge b) & -\pi(a\wedge\bar{b})\pi(\bar{a}\wedge\bar{b}) \\
-\pi(\bar{a}\wedge b)\pi(a\wedge b) & -\pi(\bar{a}\wedge b)\pi(a\wedge\bar{b}) & \pi(\bar{a}\wedge b)[1-\pi(\bar{a}\wedge b)] & -\pi(\bar{a}\wedge b)\pi(\bar{a}\wedge\bar{b}) \\
-\pi(\bar{a}\wedge\bar{b})\pi(a\wedge b) & -\pi(\bar{a}\wedge\bar{b})\pi(a\wedge\bar{b}) & -\pi(\bar{a}\wedge\bar{b})\pi(\bar{a}\wedge b) & \pi(\bar{a}\wedge\bar{b})[1-\pi(\bar{a}\wedge\bar{b})]
\end{array}\right] \qquad (22) $$

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

Nous avons déjà exprimé les indices ${}_1r(a,b)$ et ${}_3r(a,b)$ (cf. (4) et (5)) calculés au niveau de l'échantillon E, ainsi d'ailleurs que ceux correspondants ${}_1\rho(a,b)$ et ${}_3\rho(a,b)$ (cf. (7) et (18)) définis au niveau de la population parente $\mathcal{P}$. Il nous reste à introduire les v.a. ${}_1R(a,b)$ et ${}_3R(a,b)$ dont les expressions sont respectivement identiques à celles de ${}_1r(a,b)$ et ${}_3r(a,b)$, à cela près qu'on remplace les fréquences relatives observées $p(a\wedge b)$, $p(a\wedge\bar{b})$, $p(\bar{a}\wedge b)$ et $p(\bar{a}\wedge\bar{b})$ par, respectivement, $F(a\wedge b)$, $F(a\wedge\bar{b})$, $F(\bar{a}\wedge b)$ et $F(\bar{a}\wedge\bar{b})$.

III.2.2.2. Distributions asymptotiques de ${}_1R(a,b)$ et de ${}_3R(a,b)$ dans le cadre du modèle multinomial

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

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

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

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

Pour achever l'établissement de cette propriété, il reste à montrer, d'abord pour ε=1 puis pour ε=3, que l'un au moins des quatre nombres ${}_\varepsilon\rho^o_{11}$, ${}_\varepsilon\rho^o_{10}$, ${}_\varepsilon\rho^o_{01}$ et ${}_\varepsilon\rho^o_{00}$, est différent de zéro. Pour simplifier, nous allons prendre le cas dont l'intérêt est le plus général où chacune des proportions théoriques $\pi_{11},\pi_{10},\pi_{01}$ et π_{00} est différente de zéro.

ε=1

Pour alléger les écritures, nous allons ici noter par γ l'indice ${}_1\rho$ et par, respectivement, σ,δ,η et θ les proportions $\pi_{11},\pi_{10},\pi_{01}$ et π_{00} ; enfin, nous désignerons par ϕ l'expression $\pi(a)\pi(\bar a)\pi(b)\pi(\bar b)=(\pi_{11}+\pi_{10})(\pi_{01}+\pi_{00})(\pi_{11}+\pi_{01})(\pi_{10}+\pi_{00})$

$$=(\sigma+\delta)(\eta+\theta)(\sigma+\eta)(\delta+\theta)=\phi\ .\quad(23)$$

Calcul effectué, on obtient les identités suivantes :

$$\left.\begin{array}{l} \dfrac{\partial\gamma}{\partial\sigma} = \dfrac{\delta}{\sqrt{\phi}} - \dfrac{\gamma}{2} \times \dfrac{[\pi(a)+\pi(b)]}{\pi(a)\pi(b)} \quad , \quad \dfrac{\partial\gamma}{\partial\theta} = \dfrac{\sigma}{\sqrt{\phi}} - \dfrac{\gamma}{2} \times \dfrac{[\pi(\bar{a})+\pi(\bar{b})]}{\pi(\bar{a})\pi(\bar{b})} \\ \dfrac{\partial\gamma}{\partial\delta} = -\dfrac{\eta}{\sqrt{\phi}} - \dfrac{\gamma}{2} \times \dfrac{[\pi(a)+\pi(\bar{b})]}{\pi(a)\pi(\bar{b})} \quad , \quad \dfrac{\partial\gamma}{\partial\eta} = -\dfrac{\delta}{\sqrt{\phi}} - \dfrac{\gamma}{2} \times \dfrac{[\pi(\bar{a})+\pi(b)]}{\pi(\bar{a})\pi(b)} \end{array}\right\} \quad (24)$$

En raisonnant par rapport au signe de γ, on obtient dans tous les cas de figure que deux au moins de ces quantités sont différentes de zéro ; en effet, si γ est négatif, ce sont -au moins- les deux premières, si γ est nul, ce sont les quatre nombres et si γ est positif, ce sont -au moins- les deux dernières.

$\boxed{\varepsilon=3.}$

Les notations sont les mêmes que ci-dessus, à cela près que λ désignera le coefficient ${}_3\rho$. Calcul effectué, on a les relations suivantes :

$$\left.\begin{array}{l} \dfrac{\partial\lambda}{\partial\sigma} = \{\theta\sqrt{\pi(a)\pi(b)} - \dfrac{\lambda}{2} \times [\pi(a)+\pi(b)]\}/\pi(a)\pi(b) \\ \dfrac{\partial\lambda}{\partial\theta} = \sigma/\sqrt{\pi(a)\pi(b)} \\ \dfrac{\partial\lambda}{\partial\delta} = -\{\eta\sqrt{\pi(a)\pi(b)} + \dfrac{\lambda}{2}\pi(b)\}/\pi(a)\pi(b), \quad \dfrac{\partial\lambda}{\partial\eta} = -\{\delta\sqrt{\pi(a)\pi(b)} + \dfrac{\lambda}{2}\pi(a)\}/\pi(a)\pi(b) \end{array}\right\} \quad (25)$$

Compte tenu de la condition générale que nous avons posée ($\sigma x\delta x\eta x\theta \neq 0$), on a de toute façon

$$\frac{\partial\lambda}{\partial\theta} \neq 0$$

La propriété est donc bien établie. Mais il reste et c'est d'importance si on veut avoir une idée tant soit peu opérationnelle de la précision de l'estimateur ${}_\varepsilon R(a,b)$ (ε=1 ou 3) sans biais de ${}_\varepsilon\rho(a,b)$ (ε=1 ou 3), à donner une formule plus directe et plus explicite de la variance de cet estimateur.

Pour simplifier, nous désignons dans le corollaire suivant par $\rho(a,b)$ l'un des deux indices ${}_1\rho(a,b)$ ou ${}_3\rho(a,b)$ et par $R(a,b)$, l'estimateur correspondant -qui est en fait du maximum de vraisemblance- ${}_1R(a,b)$ ou ${}_3R(a,b)$. Ce corollaire est en fait une simple remarque formelle importante pour la suite des calculs explicitant la variance de $R(a,b)$.

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

En effet, considérons le vecteur aléatoire (S_i, U_i, V_i, T_i) -associé à l'extraction du i-ème individu de l'urne 4-nomiale- dont la suite des valeurs

possibles est $((1,0,0,0),(0,1,0,0),(0,0,1,0),(0,0,0,1))$ où la suite des probabilités respectivement affectées est $(\pi_{11},\pi_{10},\pi_{01},\pi_{00})$.

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

$$\rho^o_{11}\, S_i + \rho^o_{10}\, U_i + \rho^o_{01}\, V_i + \rho^o_{11}\, T_i, \qquad (26)$$

peut se mettre sous la forme :

$$(\rho^o_{11},\rho^o_{10},\rho^o_{01},\rho^o_{00})\,\mathcal{E}\{{}^t(S_i-\pi_{11},U_i-\pi_{10},V_i-\pi_{01},T_i-\pi_{00})(S_i-\pi_{11},U_i-\pi_{10},V_i-\pi_{01},T_i-\pi_{00})\}{}^t(\rho^o_{11},\rho^o_{10},\rho^o_{01},\rho^o_{00}) = {}^t\rho^o W \rho^o \qquad (27)$$

où ${}^t\rho^o = {}^t(\rho^o_{11},\rho^o_{10},\rho^o_{01},\rho^o_{00})$.

Finalement, on a

$$\mathrm{var}(R(a,b)) = \frac{1}{n}\{[\pi_{11}(\rho^o_{11})^2+\pi_{10}(\rho^o_{10})^2+\pi_{01}(\rho^o_{01})^2+\pi_{00}(\rho^o_{00})^2]$$

$$-(\pi_{11}\,\rho^o_{11}+\pi_{10}\,\rho^o_{10}+\pi_{01}\rho^o_{01}+\pi_{00}\,\rho^o_{00})^2\} \qquad (28)$$

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

III.2.2.3. Moyenne et Variance de ${}_1R(a,b)$

L'objet de ce paragraphe est, comme nous le signalions ci-dessous, de toucher de plus près les valeurs de la moyenne et de la variance de ${}_1R(a,b)$. On reprend ici les notations définies ci-dessus pour $\varepsilon=1$.

THEOREME 1. L'espérance mathématique de ${}_1R(a,b)$ est nulle.

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

$$\left.\begin{aligned}
\frac{\partial\gamma}{\partial\sigma} &\quad \frac{1}{2\sqrt{\phi}}\{-\frac{[\pi(a)\pi(\bar b)+\pi(\bar a)\pi(b)]}{\pi(a)\pi(b)}x\pi(a\wedge b)+[\pi(\bar a)+\pi(\bar b)]\}\\
\frac{\partial\gamma}{\partial\theta} &\quad \frac{1}{2\sqrt{\phi}}\{-\frac{[\pi(\bar a)\pi(b)+\pi(a)\pi(\bar b)]}{\pi(\bar a)\pi(\bar b)}x\pi(\bar a\wedge\bar b)+[\pi(a)+\pi(b)]\}\\
\frac{\partial\gamma}{\partial\delta} &\quad \frac{1}{2\sqrt{\phi}}\{\frac{[\pi(a)\pi(b)+\pi(\bar a)\pi(\bar b)]}{\pi(a)\pi(\bar b)}x\pi(a\wedge\bar b)+[\pi(\bar a)+\pi(b)]\}\\
\frac{\partial\gamma}{\partial\eta} &\quad \frac{1}{2\sqrt{\phi}}\{\frac{[\pi(\bar a)\pi(\bar b)+\pi(a)\pi(b)]}{\pi(\bar a)\pi(\bar b)}x\pi(\bar a\wedge b)+[\pi(a)+\pi(\bar b)]\}
\end{aligned}\right\} \qquad (29)$$

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

$$\left.\begin{aligned}
&\{-\frac{[\pi(a)\pi(\bar{b})+\pi(\bar{a})\pi(b)]}{\pi(a)\pi(b)} \times \pi^2(a\wedge b)+[\pi(a)+\pi(\bar{b})]\pi(a\wedge b)]\}\\
&\{+\frac{|\pi(\bar{a})\pi(b)+\pi(a)\pi(\bar{b})|}{\pi(\bar{a})\pi(\bar{b})} \times \pi^2(\bar{a}\wedge\bar{b})+[\pi(a)+\pi(b)]\pi(\bar{a}\wedge\bar{b})|\}\\
&\{+\frac{[\pi(a)\pi(b)+\pi(\bar{a})\pi(\bar{b})]}{\pi(a)\pi(\bar{b})} \times \pi^2(a\wedge\bar{b})-[\pi(\bar{a})+\pi(b)]\pi(a\wedge\bar{b})|\}\\
&\{+\frac{[\pi(\bar{a})\pi(\bar{b})+\pi(a)\pi(b)]}{\pi(\bar{a})\pi(b)} \times \pi^2(\bar{a}\wedge b)-[\pi(a)+\pi(\bar{b})]\pi(\bar{a}\wedge b)|\}
\end{aligned}\right\} \quad (30)$$

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

$$\left.\begin{aligned}
&\pi(\bar{a})=1-\pi(a),\ \pi(\bar{b})=1-\pi(b)\\
&\pi(a\wedge\bar{b})=\pi(a)-\pi(a\ b),\ \pi(\bar{a}\wedge b)=\pi(b)-\pi(a\wedge b)\\
\text{et}\quad &\pi(\bar{a}\wedge\bar{b})=\pi(a\wedge b)-\pi(a)-\pi(b)+1
\end{aligned}\right\} \quad (31)$$

Mais, au préalable, le calcul va s'organiser en établissant les contributions respectives de la première puis de la deuxième colonne de la disposition (30). La contribution de la première colonne est, au facteur $1/\pi(a)\pi(\bar{a})\pi(b)\pi(\bar{b})$ près, égale à

$$[\pi(a)\pi(b)+\pi(\bar{a})\pi(\bar{b})]\{\pi(\bar{a})\pi(b)\pi^2(a\wedge\bar{b})+\pi(a)\pi(\bar{b})\pi^2(\bar{a}\wedge b)\}$$
$$-[\pi(a)\pi(\bar{b})+\pi(\bar{a})\pi(b)]\{\pi(\bar{a})\pi(\bar{b})\pi^2(a\wedge b)+\pi(a)\pi(b)\pi^2(\bar{a}\wedge\bar{b})\} \quad (32)$$

Celle de la deuxième colonne de la disposition (30), est égale à

$$\pi(a\wedge b)[\pi(\bar{a})+\pi(\bar{b})]+\pi(\bar{a}\wedge\bar{b})[\pi(a)+\pi(b)]-\pi(a\wedge\bar{b})[\pi(\bar{a})+\pi(b)]-\pi(\bar{a}\wedge b)[\pi(a)+\pi(\bar{b})]$$
$$=4[\pi(a\wedge b)-\pi(a)\pi(b)], \quad (33)$$

en tenant compte des relations (31).

Toujours en utilisant les relations (31), on développe l'intérieur des accolades de l'expression (32) par rapport à $\pi(a\wedge b)$ et en simplifiant après un minutieux calcul, on obtient pour l'expression (32)

$$-4\pi(a)\pi(\bar{a})\pi(b)\pi(\bar{b})[\pi(a\wedge b)-\pi(a)\pi(b)], \quad (34)$$

ce qui achève d'établir le résultat annoncé.

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

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

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

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

et

$\psi=\sum\{K(\alpha,\beta)/(\alpha,\beta)\in AxB\}$. (35)

Avec ces notations, on a

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

$\Phi+\frac{\gamma^2}{4\phi}\{\psi-4\left[\pi(a)\pi(\bar{a})+\pi(b)\pi(\bar{b})\right]\}$, (36) où rappelons-le, γ représente l'indice ${}_1\rho(a,b)$ et ϕ, le produit $\pi(a)\pi(\bar{a})\pi(b)\pi(\bar{b})$.

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

$$\pi(a\wedge b)\left(\frac{\partial\gamma}{\partial\sigma}\right)^2+\pi(a\wedge\bar{b})\left(\frac{\partial\gamma}{\partial\delta}\right)^2+\pi(\bar{a}\wedge b)\left(\frac{\partial\gamma}{\partial\eta}\right)^2+\pi(\bar{a}\wedge\bar{b})\left(\frac{\partial\gamma}{\partial\theta}\right)^2. \quad (37)$$

En nous reportant aux expressions (24) ci-dessus, on a :

$$\left.\begin{array}{llll}
\left(\frac{\partial\gamma}{\partial\gamma}\right)^2=\frac{\pi^2(\bar{a}\wedge\bar{b})}{\phi}+\gamma^2\times\frac{\left[\pi(a)+\pi(b)\right]^2}{4\pi^2(a)\pi^2(b)}-\frac{\gamma}{\sqrt{\phi}}\times\frac{\left[\pi(a)+\pi(b)\right]}{\pi(a)\pi(b)}\times\pi(\bar{a}\wedge\bar{b}) & \pi(a\wedge b)\\
\left(\frac{\partial\gamma}{\partial\theta}\right)^2=\frac{\pi^2(a\wedge b)}{\phi}+\gamma^2\times\frac{\left[\pi(\bar{a})+(\bar{b})\right]^2}{4\pi^2(\bar{a})\pi^2(\bar{b})}-\frac{\gamma}{\sqrt{\phi}}\times\frac{\left[\pi(\bar{a})+\pi(\bar{b})\right]}{\pi(\bar{a})\pi(\bar{b})}\times\pi(a\wedge b) & \pi(\bar{a}\wedge\bar{b})\\
\left(\frac{\partial\gamma}{\partial\delta}\right)^2=\frac{\pi^2(\bar{a}\wedge b)}{\phi}+\gamma^2\times\frac{\left[\pi(a)+\pi(\bar{b})\right]^2}{4\pi^2(a)\pi^2(\bar{b})}-\frac{\gamma}{\sqrt{\phi}}\times\frac{\left[\pi(a)+\pi(\bar{b})\right]}{\pi(a)\pi(\bar{b})}\times\pi(\bar{a}\wedge b) & \pi(a\wedge\bar{b})\\
\left(\frac{\partial\gamma}{\partial\theta}\right)^2=\frac{\pi^2(a\wedge\bar{b})}{\phi}+\gamma^2\times\frac{\left[\pi(\bar{a})+\pi(b)\right]}{4\pi^2(\bar{a})\pi^2(b)}-\frac{\gamma}{\sqrt{\phi}}\times\frac{\left[\pi(\bar{a})+\pi(b)\right]}{\pi(\bar{a})\pi(b)}\times\pi(a\wedge\bar{b}) & \pi(\bar{a}\wedge b)
\end{array}\right\}$$

① ② ③ (38)

Relativement à la disposition (38) des calculs, on voit clairement que la contribution de la colonne ① à l'expression (37) est égale à Φ, que celle de la colonne ② est égale à $(\gamma^2\psi/4\phi)$. Enfin, la contribution de la colonne 3 nécessite calcul et simplification la ramenant à la forme $(-\gamma^2\left[\pi(a)\pi(\bar{a})+\pi(b)\pi(\bar{b})\right]/\phi)$.

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

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

On remarquera que dans le cas -qui n'a pas de sens réel- où l'attribut b est identique à l'attribut a, la variance de ${}_1R(a,b)$ est nécessairement nulle puisque R(a,b) est identiquement égal à 1. Pour des raisons de cohérence, on

le vérifie d'ailleurs aisément sur l'expression (36) ci-dessus où alors $\Phi=1/\pi(a)\pi(a)$, $\gamma^2=1$, $\phi=\pi^2(a)\pi^2(a)$ et $\psi=4\pi(a)\pi(a)$.

Pour terminer, nous allons illustrer la valeur de var.$({}_1R(a,b))$ donné par l'expression (36), dans quelques situations particulières :

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

On se rend compte, mais il y a lieu de le préciser par une tabulation plus importante, que comme on peut s'y attendre intuitivement (cf. §II ci-dessus), la variance faiblit dans le cas de forte liaison (voir ligne 1 du tableau ci-dessus et surtout et surtout ligne 4 qui correspond à une inclusion totale $E(a)\subset E(b)$) et semble avoisiner la valeur (1/n) autour de l'indépendance. De toute façon, on peut donner une valeur approximative de (36) en remplaçant les proportions théoriques par celles estimées au niveau de l'échantillon E.

III.2.2.4. Moyenne et Variance de ${}_3R(a,b)$

Nous reprenons ici la notation λ pour le coefficient ${}_3\rho$, comme c'était le cas au paragraphe III.2.2.2.

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

$$\lambda(a,b) \text{ et } \frac{1}{\eta\pi(a)\pi(b)}\{\pi(a\wedge b)\pi(\bar{a}\wedge\bar{b})\left[\pi(a\wedge b)+\pi(\bar{a}\wedge\bar{b})\right]+\pi(a\wedge\bar{b})\pi(\bar{a}\wedge b)\left[\pi(a\wedge\bar{b})+\pi(\bar{a}\wedge b)\right]-\frac{1}{4}\lambda^2(a,b)\left[\left[\pi(a\wedge\bar{b})+\pi(\bar{a}\wedge b)\right]+2\left[\pi(a)+\pi(b)+2\pi(a)\pi(b)\right]\right]\} \qquad (39)$$

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

$$\left.\begin{array}{ll} \dfrac{\partial\lambda}{\partial\sigma}=\dfrac{\pi(\bar{a}\wedge\bar{b})}{\sqrt{\pi(a)\pi(b)}}-\lambda\dfrac{\left[\pi(a)+\pi(b)\right]}{2\pi(a)\pi(b)} & \dfrac{\partial\lambda}{\partial\delta}=\dfrac{-\pi(\bar{a}\wedge b)}{\sqrt{\pi(a)\,\pi(b)}}-\lambda\times\dfrac{\pi(b)}{2\pi(a)\pi(b)} \\ \dfrac{\partial\lambda}{\partial\theta}=\dfrac{\pi(a\wedge b)}{\sqrt{\pi(a)\pi(b)}} & \dfrac{\partial\lambda}{\partial\eta}\quad\dfrac{-\pi(a\wedge\bar{b})}{\sqrt{\pi(a)\pi(b)}}-\lambda*\dfrac{\pi(a)}{2\pi(a)\pi(b)} \end{array}\right\} \qquad (40)$$

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

Il reste maintenant à déterminer la moyenne des carrés :

$$\pi(a\wedge b)(\partial\lambda/\partial\sigma)^2+\pi(a\wedge\bar{b})(\partial\lambda/\partial\delta)^2+\pi(\bar{a}\wedge b)(\partial\lambda/\partial\eta)^2+\pi(\bar{a}\wedge\bar{b})(\partial\lambda/\partial\theta)^2 \qquad (42)$$

$$\left(\frac{\partial\lambda}{\partial\sigma}\right)^2 = \frac{1}{\pi(a)\pi(b)}\left\{\pi^2(\bar{a}\wedge\bar{b}) - \frac{[\pi(a)+\pi(b)]\,\pi(\bar{a}\wedge\bar{b})}{\sqrt{\pi(a)\pi(b)}} \times \lambda + \frac{[\pi(a)+\pi(b)]^2}{4\,\pi(a)\pi(b)}\lambda^2\right\} \pi(a\wedge b)$$

$$\left(\frac{\partial\lambda}{\partial\delta}\right)^2 = \frac{1}{\pi(a)\pi(b)}\left\{\pi^2(\bar{a}\wedge b) + \frac{\pi(b)\pi(\bar{a}\wedge b)}{\sqrt{\pi(a)\pi(b)}} \times \lambda + \frac{[\pi(b)]^2}{4\,\pi(a)\pi(b)}\lambda^2\right\} \pi(a\wedge\bar{b})$$

$$\left(\frac{\partial\lambda}{\partial\eta}\right)^2 = \frac{1}{\pi(a)\pi(b)}\left\{\pi^2(a\wedge\bar{b}) + \frac{\pi(a)\pi(a\wedge\bar{b})}{\sqrt{\pi(a)\pi(b)}} \times \lambda + \frac{[\pi(a)]^2}{4\,\pi(a)\pi(b)}\right\} \pi(\bar{a}\wedge b)$$

$$\left(\frac{\partial\lambda}{\partial\theta}\right) = \frac{\pi^2(a\wedge b)}{\pi(a)\pi(b)} \qquad \pi(\bar{a}\wedge\bar{b})$$

① ② ③ (43)

La contribution de la colonne ① est

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

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

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

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

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

D'où, la variance des nombres de (40) relativement à $\{\pi(\alpha,\beta)/(\alpha,\beta)\in AxB\}$:

$$\frac{1}{\pi(a)\pi(b)}\{\pi(a\wedge b)\pi(\bar{a}\wedge\bar{b})[\pi(a\wedge b)+\pi(\bar{a}\wedge\bar{b})] + \pi(a\wedge\bar{b})\pi(\bar{a}\wedge b)[\pi(a\wedge\bar{b})+\pi(\bar{a}\wedge b)]$$
$$-\frac{\lambda^2}{4}[[\pi(a\wedge\bar{b})+\pi(\bar{a}\wedge b)]+2[\pi(a)+\pi(b)+2\pi(a)\pi(b)]]\}, \qquad (44)$$

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

Nous allons nous contenter d'illustrer numériquement une fois l'expression (44).

Considérons l'exemple suivant : $\pi(a\wedge b)=0,4$, $\pi(a\wedge\bar{b})=0,1$, $\pi(\bar{a}\wedge b)=0,1$ et $\pi(\bar{a}\wedge\bar{b})=0,4$. On obtient la valeur 0,232 pour l'expression (44).

III.2.2.5. Intervalle de confiance pour ${}_1\rho(a,b)$ et pour ${}_3\rho(a,b)$; conséquence sur la précision calcul nécessaire en analyse des données.

C'est pour ne pas avoir à traîner des pré-indices que nous avons ci-dessus noté γ le coefficient ${}_1\rho$ et λ, celui ${}_3\rho$. Désignons par C^2/n et par L^2/n les variances respectives de ${}_1R(a,b)$ et de ${}_3R(a,b)$ (cf. les expressions (36) et (39)).

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

$$\left[{}_1r(a,b)-\sqrt{C^2/n}\;\; G^{-1}(1-\tfrac{\alpha}{2}),\; {}_1r(a,b)+\sqrt{C^2/n}\;\; G^{-1}(1-\tfrac{\alpha}{2})\right] \qquad (45)$$

et

$$\left[{}_3r(a,b)-\sqrt{L^2/n}\;\; G^{-1}(1-\tfrac{\alpha}{2}),\; {}_3r(a,b)+\sqrt{L^2/n}\;\; G^{-1}(1-\tfrac{\alpha}{2})\right], \qquad (46)$$

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

En considérant un seuil de confiance de l'ordre de 0,99, les intervalles de confiance (45) et (46) deviennent, lorsque a et b sont indépendants (au niveau de la population parente) :

$$\left[{}_1r(a,b)-2,5/\sqrt{n},\;\; {}_1r(a,b)+2,5/\sqrt{n}\;\right] \qquad (47)$$

$$\left[{}_3r(a,b)-2,5\sqrt{\pi(\bar{a})\pi(\bar{b})}/\sqrt{n},\;\; {}_3r(a,b)+2,5\sqrt{\pi(\bar{a})\pi(\bar{b})}/\sqrt{n}\;\right] ; \qquad (48)$$

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

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

Dans ces conditions, dans le calcul de l'indice d'association entre variables au niveau d'un échantillon, il n'est pas de sens d'aller au delà d'un certain nombre de chiffres significatifs, lequel nombre dépendant de l'effectif de l'échantillon ; d'où l'aberration des programmes d'analyse des données qui travaillent avec la double précision, alors que l'effectif -d'ailleurs important de l'échantillon- ne permet en fait que deux ou trois chiffres significatifs pour l'estimation des corrélations entre variables...

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

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

pu tester que la qualité de ces méthodes restait très intimement liée à celle de l'indice de ressemblance ayant permis d'établir la préordonnance.

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

$$\Pr\{{}_1R(a,b)<{}_1R(c,d)\}=G(8/\sqrt{2})\cong 1, \quad (49)$$

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

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

Compte tenu de ce dernier fait, mais en pondérant également par la qualité de la précision des intervalles de confiance, le nombre k de chiffres significatifs après la virgule que nous préconisons pour la mesure des coefficients d'association entre variables, est défini par $10^k\leqslant n<10^{(k+1)}$!

III.2.3. Comparaison entre deux associations respectivement relatives à deux paires d'attributs observées sur deux échantillons de tailles distinctes; un illogisme statistique ?

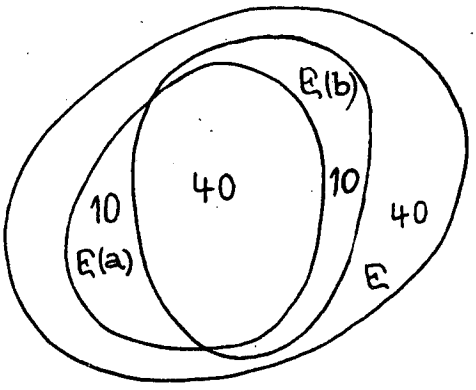
Pour des raisons de simplicité d'interprétation, les deux paires d'attributs {a,b} et {c,d} sont supposées sans composante commune. D'autre part, on considèrera ici l'indice d'expression plus simple $_3\rho(u,v)$ défini au niveau de la population entière, celui correspondant $_3r(u,v)$ défini au niveau de l'échantillon E et enfin celui toujours défini au niveau de E, mais basé sur la vraisemblance de la relation $\Phi\left[\sqrt{n}\ {}_3r(u,v)\right]$ (cf. formule (17) § III.1.), où Φ est la f.r. de la loi normale centrée réduite.

Par rapport à une optique ascendante où la taille n de l'échantillon E augmente, on ne peut se servir de ce dernier indice pour comparer les associations entre a et b d'une part, c et d d'autre part. En effet, pourvu que $_3\rho(a,b)$ et $_3\rho(c,d)$ soient non négligeables et de même signe -positif par exemple-, avec une quasi certitude $\Phi\left[\sqrt{n}\ {}_3r_n(a,b)\right]$ et $\Phi\left[\sqrt{n}\ {}_3r_n(c,d)\right]$ tendent très vite vers l'unité et deviennent indistinguables. Ce problème se présente bien entendu également lorsqu'il s'agit de comparer les associations deux à deux d'un ensemble $\mathcal{A}$ d'attributs descriptifs. C'est au paragraphe suivant que nous nous réservons de montrer comment utiliser de façon cohérente et justifiée d'un point de vue statistique, toute la discrimination définie par une échelle $[0,1]$ de probabilité (ou fréquence mathématique) associée à la f.r. de la loi normale centrée et réduite, pour l'évaluation des associations au moyen d'un indice basé sur la vraisemblance des liaisons observées.

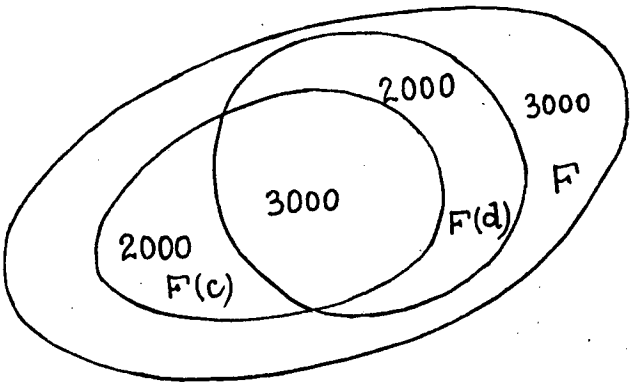
Dans le cadre de ce paragraphe (III.2.3.), la comparaison des associations entre a et b d'une part, c et d d'autre part, se fera sur la base de l'indice $_3\rho$. Il s'agit donc de comparer $_3\rho(a,b)$ et $_3\rho(c,d)$ dont les estimations de maximum de vraisemblance sur deux échantillons aléatoires indépendants de tailles respectives n et m, sont notées $_3r_n(a,b)$ et $_3r_m(a,b)$.

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

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



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



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

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

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

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

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

Si on considère maintenant dans chacun des deux cas le test de l'hypothèse d'indépendance entre les deux attributs au seuil $\alpha= 0,001$ (cf. § III.2.1.). On se trouve conduit à ne pas rejeter l'hypothèse d'indépendance pour la relation entre a et b, mais par contre, à rejeter violemment une telle hypothèse d'indépendance pour l'association entre c et d !

Mais alors imaginons que la situation réelle en ce qui concerne la valeur de ${}_3\rho(c,d)$ soit celle la plus vraisemblable ; c'est-à-dire ${}_3\rho(c,d)=0,1$. D'ailleurs, le calcul de l'intervalle de confiance à 0,99 pour ${}_3\rho(a,b)$ donne $[0,088;0,112]$ (voir pour ce calcul l'expression (44) ci-dessus).

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

$$\Pr\{{}_3R_n(a,b)\geqslant 0,3 / {}_3\rho(a,b)\leqslant 0,1\}\leqslant \Pr\{\frac{{}_3R_n(a,b)-0,1}{\sqrt{\mathrm{var}[{}_3R_n(a,b)]}}\geqslant \frac{0,2}{\sqrt{0,232/100}}\cong 4,15\}\cong 10^{-5}, \qquad (50)$$

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

Ainsi, dans la situation la plus vraisemblable (${}_3\rho(c,d)=0,1$), une hypothèse telle que ${}_3\rho(a,b)\leqslant {}_3\rho(c,d)$ apparaît comme hautement invraisemblable (cf. (50)) et pourtant -répétons-le- la pratique ci-dessus du test d'indépendance au seuil $\alpha=0,001$, nous fait violemment rejeter l'hypothèse d'indépendance pour la relation entre c et d, mais, ne nous fait pas exclure une telle indépendance entre a et b ! D'où le paradoxe.

Cette dernière circonstance nous renforce dans notre démarche qui consiste à utiliser l'h.a.l. "latérale" (définie de façon plus ou moins combinatoire au niveau de E) pour la construction d'indices normalisés et pour la détermination d'une échelle $[0,1]$ de fréquence mathématique -destinée à la comparaison deux à deux d'un ensemble de variables descriptives (ici formée d'attributs)- en termes de vraisemblance des liaisons observées. Nous l'avons déjà dit ci-dessus (comparaison de deux paires d'attributs) et nous le répéterons ci-dessous, une telle échelle ne peut être directement établie à partir d'une formule telle que (6) ci-dessus.

IV. ECHELLE DE PROXIMITE POUR LA COMPARAISON DEUX A DEUX D'UN ENSEMBLE DE VARIABLES

IV.1. Introduction ; la pratique de la réduction globale des indices d'association

Relativement au problème de l'organisation en classes et sous-classes de proximité d'un ensemble $\mathcal{A}$ de variables descriptives (il s'agit ici le plus souvent d'attributs), la structure de proximité nécessaire à l'application de l'algorithme de la vraisemblance du lien [LERMAN(1970) à voir dans (1981) chap. 5], se trouve définie par la table (indexée par l'ensemble des paires de $\mathcal{A}$) :

$$\{P(a,b)/\{a,b\}\in P_2(\mathcal{A})\}, \qquad (1)$$

où, si (a',b') est un couple de v.a. indépendantes, de même type et respectant les caractéristiques cardinales de (a,b), alors P(a',b') est une variable aléatoire uniforme sur l'intervalle $[0,1]$. C'est "théoriquement" le cas pour l'indice défini par l'expression (2) (§ III.1. ci-dessus) qu'on calculera au moyen de l'expression (6) (§ III.1.) : $P(a,b)=\Phi[\sqrt{(n-1)}\ r(a,b)]$ où Φ est la f.r. de la loi normale $N(0,1)$; en effet $P(a,b)=\Pr\{S\leqslant s=n(a\wedge b)/N_1\}$ apparaît comme la valeur de la fonction de répartition de $n(a'\wedge b')$.

La première analyse du comportement de l'algorithme de la vraisemblance du lien a été effectué par Mme Nicolau [M.H. NICOLAU(1972)] sur une famille de 110 attributs de description observés sur un ensemble de 1500 personnages

enfantins. L'usage direct des indices de la table (1) conformément à leur calcul au moyen de l'expression (6) (§ III.1.)- $P(a,b)=\Phi[\sqrt{(n-1)}\ r(a,b)]$ - conduit à un arbre des classifications très tassé où seuls quelques niveaux (trois ou quatre) très synthétiques apparaissent. Bien que les partitions obtenues soient très significatives, le but -qui consistait à utiliser une échelle d'évaluation des proximités basée sur la vraisemblance des liaisons, suffisamment fine et riche pour la discrimination- n'était pas atteint. La raison est qu'une part suffisamment appréciable pour déterminer l'allure de l'arbre des classifications, des valeurs de $Q(a,b)=\sqrt{n-1}\ r(a,b)$ (cf. (1)§III.1.) atteignait des nombres de l'ordre 3 ou 4. Encore une fois, le passage direct de $Q(a,b)$ à l'échelle de fréquence mathématique $[0,1]$, au moyen de la transformation monotone $\Phi[Q(a,b)]$, permet davantage la mise en évidence d'un lien fort entre les deux composantes d'une même paire d'attributs que l'évaluation comparée des liaisons entre les différentes paires d'un même ensemble $\mathcal{A}$ d'attributs ou de variables descriptives.

La première tentative de réduction globale des similarités $\{Q(a,b)/\{a,b\} \in P_2(\mathcal{A})\}$ répondait au souci d'utiliser tout le pouvoir discriminant de l'échelle $[0,1]$ de probabilité de la f.r. de la loi normale $\mathcal{N}(0,1)$ [M.H. NICOLAU(1972), I.C. LERMAN(1973)]. A cette fin, on définit un paramètre de réduction λ positif tel que

$$\max\{Q(a,b)/\{a,b\} \in P_2(\mathcal{A})\}/\lambda = 2,5$$

soit

$$\lambda = 2,5/\max\{Q(a,b)/\{a,b\} \in P_2(\mathcal{A})\} \qquad (2)$$

On substitue alors à la première table des indices $\{Q(a,b)/\{a,b\} \in P_2(\mathcal{A})\}$, la table des indices réduits au moyen du paramètre λ

$$\{Q_\lambda(a,b)/\{a,b\} \in P_2(\mathcal{A})\}, \quad (3)$$

où $Q_\lambda(a,b)=Q(a,b)/\lambda$ pour tout $\{a,b\} \in P_2(\mathcal{A})$.

La table définitive des indices qui se réfèrent à une échelle $[0,1]$ de probabilité :

$$\{P_\lambda(a,b)/\{a,b\} \in P_2(\mathcal{A})\} \qquad (4)$$

est à ce moment obtenue au moyen de la transformation

$$P_\lambda(a,b)=\Phi[Q_\lambda(a,b)] \qquad (5)$$

pour tout $\{a,b\} \in P_2(\mathcal{A})$, où Φ est toujours la f.r. de la loi $(0,1)$.

L'algorithme de la vraisemblance du lien a, en partant de la table (4) des indices de proximité, pu conduire aux meilleurs résultats dans de nombreuses études [cf.LERMAN(1981), Partie II].

Un progrès supplémentaire a été obtenu quant à la cohérence des associations qui se produisent aux derniers niveaux de l'arbre, au moyen d'un autre type de réduction globale des similarités $\{Q(a,b)/\{a,b\} \in P_2(\mathcal{A})\}$. Si le premier mode correspond à se ramener par homothétie à une distribution dont la valeur maximale est inférieure ou égale au quantile 0.994 de la loi normale, le second opère par transformation affine pour se ramener à une distribution des similarités de moyenne nulle et de variance unité. Ainsi, à partir des $Q(a,b), \{a,b\} \in P_2(\mathcal{A})$, on définit

$$\bar{Q} = \frac{1}{\binom{m}{2}} \sum \{Q(a,b)/\{a,b\} \in P_2(\mathcal{Q})\} \qquad (6)$$

$$\mathrm{var}(Q) = \frac{1}{\binom{m}{2}} \sum \{[Q(a,b)-\bar{Q}]^2/\{a,b\} \in P_2(\mathcal{Q})\} \qquad (7)$$

La nouvelle distribution des similarités considérée avant la référence à une échelle $[0,1]$ de probabilité (via la loi normale) est définie par

$$\{Q_r(a,b) = \frac{Q(a,b)-\bar{Q}}{\sqrt{\mathrm{var}(Q)}} / \{a,b\} \in P_2(\mathcal{Q})\} \qquad (8)$$

Tout se passe comme s'il y avait lieu de définir -pour l'évaluation comparée des liens au moyen d'un indice de vraisemblance- un modèle aléatoire de l'hypothèse d'absence de liaisons ayant un caractère conditionnel en ajustant de façon globale les paramètres du modèle à ceux de la distribution observée des similarités $Q(a,b), \{a,b\} \in P_2(\mathcal{Q})$.

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

Nous avons pu constater que dans la pratique des données réelles $\bar{Q}$ avoisinait la valeur zéro. De sorte que -et d'ailleurs même indépendamment- nous envisageons ici un mode de réduction plus simple que celui (8) ci-dessus, où la table des similarités globalement normalisées se trouve définie par

$$\{Q_s(a,b) = \frac{Q(a,b)}{\sqrt{M_2(Q)}} / \{a,b\} \in P_2(\mathcal{Q})\}, \qquad (9)$$

où

$$M_2(Q) = \frac{1}{\binom{m}{2}} \sum \{Q^2(a,b)/\{a,b\} \in P_2(\mathcal{Q})\}, \qquad (10)$$

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

C'est ce dernier type de réduction global que nous chercherons à justifier au mieux d'un point de vue statistique. En des termes plus précis et en considérant le cas où $\mathcal{Q}$ est un ensemble d'attributs descriptifs, l'hypothèse d'absence de liaison N_1, N_2 ou N_3 (cf. § III.1., III.2.1.) va associer un ensemble $\mathcal{Q}'$ d'attributs aléatoires indépendants et par conséquent, la famille des indices aléatoires d'association :

$$\{Q_s(a',b')/\{a',b'\} \in P_2(\mathcal{Q}')\}, \quad (11)$$

où

$$Q_s(a',b')=Q(a',b')\Big/\left[\frac{1}{\binom{m}{2}} \sum \{Q^2(a',b')/\{a',b'\} \in P_2(\ ')\}\right]^{1/2} \quad (12)$$

Il s'agit alors de justifier au mieux que, relativement à un couple (a,b) d'attributs, on peut admettre la référence à la loi normale $\mathcal{N}(0,1)$ pour la distribution de $Q_s(a',b')$, ce qui permet avec une rigueur suffisante d'établir un indice basé sur la vraisemblance de la liaison au moyen de la formule

$$P_s(a,b)= \Phi \left[Q_s(a,b)\right] \quad (13)$$

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

Pour terminer ce paragraphe, remarquons que l'indice $Q_s(a,b)$ qui rapporte $Q(a,b)$ à $\sqrt{M_2(Q)}$ se présente sous la forme d'une "densité orientée" en $\{a,b\}$ de $P_2(\mathcal{Q})$ et qu'on a

$$\sum\{Q_s^2(a,b)/\{a,b\} \in P_2(\mathcal{Q})\} = \binom{m}{2} \quad (14)$$

IV.2. Quelques résultats relatifs à la distribution de la table des indices d'association entre v.a. "non-paramétriques" associés à une suite de variables observées

IV.2.1. Position du problème

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

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

$$\mu_j = \frac{1}{n} \sum_{1\leq i\leq n} v_j(i) \text{ et } \mathrm{var}(j)= \frac{1}{n} \sum_{1\leq i\leq n} \left[v_j(i)-\mu_j\right]^2 , \quad (15)$$

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

En posant w_j la j-ème variable centrée réduite :

$$w_j=\left[v_j-\mu_j\right]/\sqrt{\mathrm{var}(j)} \quad (16)$$

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

$$Q(w_j.\sigma_j,w_k.\sigma_k)= \frac{1}{\sqrt{n}} \sum_{1\leq i\leq n} w_j.\sigma_j(i)\ w_k.\sigma_k(i) \quad (17)$$

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

$$\{Q(w_j.\sigma_j, w_k.\sigma_k) = \frac{1}{\sqrt{n}} \sum_{1\leq i\leq n} \{ w_j.\sigma_j(i)\ w_k.\sigma_k(i) / 1\leq j\leq k\leq m \} \qquad (18)$$

Ce problème a un caractère tout à fait fondamental puisqu'il est en quelque sorte le correspondant non paramétrique du vieux problème résolu par J. Wishart [J. WISHART(1928)] dans le cas où les v.a. suivent une loi multinormale. Les résultats que nous obtiendrons ici auront un caractère encore parcellaire, mais nous l'espérons suffisant pour justifier du bien fondé de la référence à la loi normale utilisée dans la formule (13) ci-dessus.

Nous allons commencer par déterminer (§ IV.2.2.) la matrice des covariances de la suite des v.a. $\{Q(w'_j,w'_k)/1\leq j\leq k\leq m\}$, où nous avons noté w'_j et w'_k pour $w_j.\sigma_j$ et $w_k.\sigma_k$. On verra que cette dernière a une forme très simple. Nous déterminerons ensuite (§ IV.2.3.) cette même matrice dans le cas où l'ensemble des variables est un ensemble $\mathcal{A}$ d'attributs descriptifs et ce, pour chacune des trois hypothèses d'absence de liaison N_1, N_2 et N_3. Au paragraphe IV.2.4., nous admettrons une conjecture relative à la loi de la somme des carrés $\sum\{Q^2(w'_j,w'_k)/1\leq j<k\leq m\}$. C'est en considérant la situation la plus simple de trois attributs que nous nous rendrons compte (§ IV.2.5.) du caractère très particulier de la loi de probabilité de $\{Q(a'_j,a'_k)/1\leq j<k\leq m\}$: table des indices aléatoires associée à $\mathcal{A}$. Le paragraphe IV.2.6. est sans doute le plus important puisqu'on précise une hypothèse d'absence de liaison -définie au niveau de l'ensemble $\mathcal{A}$ des attributs- compatible avec N_2, pour laquelle la distribution de la table $\{Q(a'_j,a'_k)/1\leq j<k\leq m\}$ est asymptotiquement multinormale ; pour une telle hypothèse, la réduction de l'indice Q(a,b) se fera au moyen d'une forme quadratique non identique à (10).

IV.2.2. Matrice des covariances de la table (18)

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

IV.2.2.1. Calcul de $\mathcal{E}[Q(w'_j,w'_k)Q(w'_h,w'_\ell)]$

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

$$\text{moy.}[Q(w_j.\sigma_j,w_k.\sigma_k)] = n\ \text{moy.}(w_j)\text{moy.}(w_k) = 0 \qquad (19)$$

$$\text{var.}[Q(w_j.\sigma_j,w_k.\sigma_k)] = \frac{n^2}{(n-1)}\ \text{var.}(w_j)\text{var.}(w_k) = \frac{n}{(n-1)} \cong 1 \qquad (20)$$

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

IV.2.2.2. Calcul de $\mathcal{E}\left[Q(w_j',w_k')Q(w_j',w_\ell')\right]$

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

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

$$= \frac{1}{(n!)^2}\sum\left\{\left[\sum_i w_j(i)w_k.\sigma_k(i)\right]\left[\sum_{i'} w_j(i')w_\ell.\sigma_\ell(i')\right]\Big/(\sigma_k,\sigma_\ell)\in G_n\times G_n\right\}, \quad (21)$$

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

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

$$\left(\sum_i w_j(i)\left[\frac{1}{n!}\sum\{w_k.\sigma_k(i)/\sigma_k\in G_n\}\right]\right)\left(\sum_{i'} w_j(i')\left[\frac{1}{n!}\sum\{w_\ell.\sigma_\ell(i')/\sigma_\ell\in G_n\}\right]\right) \quad (22)$$

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

$$(n\ \text{moy.}(w_j)\ \text{moy.}(w_k))\times(n\ \text{moy.}(w_j)\text{moy.}(w_\ell))=0 \quad (23)$$

Il est donc a fortiori vrai que

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

$$= (n\ \text{moy.}(w_j)\text{moy.}(w_k))\times(n\ \text{moy.}(w_j)\text{moy.}(w_\ell))=0 \quad (24)$$

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

IV.2.2.3. Calcul de $\mathcal{E}\left(\left[Q(w_j',w_k')\right]^2\right)$

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

Dans ces conditions, la matrice des covariances de la table (18) des indices aléatoires est exactement la matrice identité.

D'autre part, comme nous l'avons mentionné au paragraphe II, en vertu d'un célèbre théorème de la Statistique non paramétrique [WALD&WOLFOWITZ(1944), NOETHER(1949),HAJEK(1961)], la loi marginale de $Q(w_j.\sigma_j,w_k.\sigma_k)$ est, pour tout $(j,k):1\leqslant j<k\leqslant m$, asymptotiquement normale et centrée réduite dans notre cas (cf. (19) et (20)).

Toutefois, bien entendu, la loi jointe de (18) n'est pas normale, car si les v.a. composantes de (18) sont deux à deux indépendantes, elles ne le sont plus trois à trois comme on s'en rendra compte au paragraphe IV.2.5.

IV.2.3. Matrice des covariances de la table des indices aléatoires d'association entre attributs

Nous désignerons cette matrice par $\{Q(a'_j,a'_k)/1\leqslant j<k\leqslant m\}$ où a'_j est l'attribut aléatoire associé à a_j, $1\leqslant j\leqslant m$, dans le cadre de l'une des h.a.1. N_1, N_2 ou N_3 (cf.§ III.1. et § III.2.1.). $Q_\varepsilon(a'_j,a'_k)$ est l'indice aléatoire centré et réduit conformément à l'une des h.a.1. N_1, N_2 ou N_3 (ε=1,2 ou 3 pour N_1, N_2 ou N_3).

IV.2.3.1. h.a.1. N_1

Cette hypothèse d'absence de liaison est de même nature que celle à caractère permutationnelle considérée ci-dessus. Il suffit en effet de considérer cette dernière en interprétant un attribut comme une variable quantitative à valeurs 0 et 1, pour retrouver une équivalence avec l'h.a.1. N_1. Les résultats sont donc les mêmes que ci-dessus (§ IV.2.2.). C'est précisément relativement à la comparaison deux à deux d'un ensemble de trois attributs aléatoires {a',b',c'} que nous mettrons en évidence (§ IV.2.5.) le caractère particulier de la loi jointe des trois indices aléatoires d'association entre a',b' et c'.

IV.2.3.2. h.a.1. N_2

En se reportant au paragraphe IV.2.2. ci-dessus, on se rend compte que la seule configuration non triviale pour la comparaison -au moyen de la covariance- de deux indices aléatoires d'association, est celle où les deux paires respectives sur lesquelles portent les deux indices, ont une composante commune. Dans ces conditions désignons par X, Y et Z les trois parties définies par trois attributs aléatoires a',b' et c'. Nous allons directement déterminer l'espérance mathématique :

$$\mathcal{E}\left[\mathrm{card}(X\cap Y)\cdot\mathrm{card}(X\cap Z)\right] \qquad (25)$$

du produit des deux indices aléatoires associés aux indices bruts s(a,b) et s(a,c).

Le calcul de (25) nécessite la détermination dans l'h.a.1. N_2, de la probabilité conditionnelle

$$\Pr\{\mathrm{card}(X\cap Y)=r,\ \mathrm{card}(X\cap Z)=s,\ \mathrm{card}(Y)=\ell,\ \mathrm{card}(Z)=m/\mathrm{card}(X)=k\}, \qquad (26)$$

où $r\leqslant\min(k,\ell)$, $s\leqslant\min(k,m)$.

En désignant par α,β et γ les proportions $\mathrm{card}[E(a)]/n$, $\mathrm{card}[E(b)]/n$ et $\mathrm{card}[E(c)]/n$, où a, b et c sont les attributs observés auxquels correspondent a',b' et c', la probabilité (26) peut s'écrire

$$\frac{\binom{k}{r}\binom{n-k}{\ell-r}}{\binom{n}{\ell}} \times \frac{\binom{k}{s}\binom{n-k}{m-s}}{\binom{n}{m}} \times \binom{n}{\ell}\beta^{\ell}(\bar\beta)^{(n-\ell)}\binom{n}{m}\gamma^{m}(\bar\gamma)^{(n-m)}, \qquad (27)$$

où nous avons noté $\bar\beta$ pour $(1-\beta)$ et $\bar\gamma$ pour $(1-\gamma)$. De sorte que

$$\Pr\{\mathrm{card}(X\cap Y)=r,\ \mathrm{card}(X\cap Z)=s/\mathrm{card}(X)=k\}$$
$$=\sum\{\binom{k}{r}\binom{n-k}{\ell-r}\binom{k}{s}\binom{n-k}{m-s}\beta^{\ell}(\bar\beta)^{n-\ell}\gamma^{m}(\bar\gamma)^{n-m}/\ell\geq r,\ m\geqslant s\} \qquad (28)$$

$r \leqslant k <=> (n-r) \geqslant (n-k)$, ainsi $(\ell-r)$ atteint $(n-k)$, de même,
$s \leqslant k <=> (n-s) \geqslant (n-k)$, ainsi $(m-s)$ atteint $(n-k)$.

En effectuant le changement de variables : $p=(\ell-r)$ et $q=(m-s)$, l'expression (28) devient

$$\sum\{\binom{k}{r}\binom{k}{s}\beta^r\gamma^s\binom{n-k}{p}\beta^p(\bar{\beta})^{n-p-r}\binom{n-k}{q}\gamma^q(\bar{\gamma})^{n-q-s} / 0 \leqslant p \leqslant (n-k), 0 \leqslant q \leqslant (n-k)\} \qquad (29)$$

En écrivant $(n-p-r)$ (resp. $(n-q-s)$) sous la forme $\{[(n-k)-p]+(k-r)\}$ (resp. $\{[(n-k)-q]+(k-s)\}$), on obtient après sommation

$$[\binom{k}{r}\beta^r(\bar{\beta})^{k-r}]\,[\binom{k}{s}\gamma^s(\bar{\gamma})^{k-s}] \ , \qquad (30)$$

ce qui établit l'indépendance des deux v.a. $\mathrm{card}(X \cap Y)$ et $\mathrm{card}(X \cap Z)$, conditionnellement à la donnée de $\mathrm{card}(X)$. Maintenant, l'expression (25) peut se décomposer comme suit

$$\sum_{0 \leqslant k \leqslant n} \mathscr{P}\{\mathrm{card}(X \cap Y)\mathrm{card}(X \cap Z)/\mathrm{card}(X)=k\}\ \Pr\{\mathrm{card}(X)=k\}$$

$$= \sum_{0 \leqslant k \leqslant n} (k\beta)(k\gamma) \times \binom{n}{k}\ \alpha^k(\bar{\alpha})^{n-k} = (n\alpha\beta)(n\alpha\gamma)+n\alpha\bar{\alpha}\beta\gamma \qquad (31)$$

Finalement, on a

$$\mathrm{Cov.}(\mathrm{card}(X \cap Y), \mathrm{card}(X \cap Z)) = n\alpha\bar{\alpha}\beta\gamma \qquad (32)$$

et $$\mathrm{Cor.}(\mathrm{card}(X \cap Y), \mathrm{card}(X \cap Z)) = \bar{\alpha}\beta\gamma/\sqrt{\beta\gamma(1-\alpha\beta)(1-\alpha\gamma)} \qquad (33)$$

Considérons à présent la totalité de la matrice des covariances associée à l'ensemble $\{s(a'_h,a'_j)/1 \leqslant h<j \leqslant m\}$ des indices bruts aléatoires ; nous pourrons indiquer par X'_j la partie aléatoire de E définie par $a'_j, 1 \leqslant j \leqslant m$. Codons la paire $\{a'_h,a'_j\}$ par hj où $1 \leqslant h<j \leqslant m$. On suppose que la matrice des covariances des $s(a'_h,a'_j)$ est indexée par la suite lexicographiquement ordonnée des hj $(1 \leqslant h<j \leqslant m)$:
12,13,...,1m,23,24,...,2m,34,35,...,3m,.....,(m-2)(m-1),(m-2)m,(m-1)m.

Cette matrice est formée de zéros en dehors d'une suite de blocs carrés diagonaux qui se suivent sur la diagonale principale et dont les dimensions respectives sont $(n-1),(n-2),\ldots,2,1$; la h-ème matrice étant indexée par $\{hj/j>h\}$. Cette h-ème matrice est la matrice des covariances entre les indices aléatoires $\{s(a'_h,a'_j)/j=(h+1),(h+2),\ldots,m\}$, elle est donc définie ; en effet, en vertu de ci-dessus, pour a'_k fixé, la suite des indices aléatoire de la dernière table est une suite de v.a. indépendantes. D'où le résultat suivant :

PROPRIETE 1. La matrice des covariances de la suite des indices aléatoires $\{s(a'_h,a'_j)/1 \leqslant h<j \leqslant m\}$ est définie.

L'expression (30) ci-dessus donne la probabilité conditionnelle

$$\Pr\{\mathrm{card}(X \cap Y)=r,\ \mathrm{card}(X \cap Z)=s/\mathrm{card}(X)=k\}, \qquad (34)$$

d'où

PROPRIETE 2. La distribution de probabilité du couple $(\mathrm{card}(X\cap Y),\mathrm{card}(X\cap Z))$ est définie par

$$\Pr\{\mathrm{card}(X\cap Y)=r,\ \mathrm{card}(X\cap Z)=s\}=\sum_{0\leq k\leq n} B(k,r,\beta)B(n,k,\gamma)B(n,k,\alpha) \quad (35)$$

où $B(k,r,\beta)$, $B(k,s,\gamma)$ et $B(n,k,\alpha)$ désignent des probabilités binomiales.

IV.2.3.3. h.a.l. N_3

PROPRIETE 3. La distribution de probabilité du couple $(\mathrm{card}(X\cap Y),\mathrm{card}(X\cap Z))$ est définie par

$$\Pr\{\mathrm{card}(X\cap Y)=r,\mathrm{card}(X\cap Z)=s\}=\sum_{k\geq 0} B(k,r,\beta)B(k,s,\gamma)\frac{(n\alpha)^k}{k!}\,e^{-n\alpha} \quad (36)$$

En effet, d'après la définition même du modèle aléatoire

$$\Pr\{\mathrm{card}(X\cap Y)=r,\mathrm{card}(X\cap Z)=s\}=\sum_{\nu\geq 0}\ \sum_{0\leq k\leq \nu} B(k,r,\beta)B(k,s,\gamma)B(k,\nu,\alpha)\times\frac{n^\nu}{\nu!}\,e^{-n} \quad (37)$$

En invertissant l'ordre des sommations, on obtient

$$\sum_{k\geq 0} B(k,r,\beta)B(k,s,\gamma)\Big(\sum_{\nu\geq k}\binom{\nu}{k}\alpha^k(\bar{\alpha})^{\nu-k}\,\frac{n^\nu}{\nu!}\,e^{-n}\Big)$$

or la dernière somme sous parenthèses peut se mettre sous la forme

$$\frac{(n\alpha)^k}{k!}\,e^{-n\alpha}\Big(\sum_{\nu\geq k}\frac{(n\bar{\alpha})^{\nu-k}}{(\nu-k)!}\,.e^{-n\bar{\alpha}}\Big)$$

$$=\frac{(n\alpha)^k}{k!}\,e^{-n\alpha}$$

car -en posant $\ell=\nu-k$- on voit que le contenu de la dernière parenthèse est une somme complète de probabilités de Poisson. D'où le résultat annoncé.

Dans ces conditions, déterminons la covariance entre les deux v.a. $R=\mathrm{card}(X\cap Y)$ et $S=\mathrm{card}(X\cap Z)$.

$$\mathcal{E}(R.S)=\sum_{k\geq 0}\Big[\sum_{r\leq k} rB(k,r,\beta)\Big]\Big[\sum_{s\leq k} sB(k,s,\gamma)\Big]\frac{(nd)^k}{k!}\,e^{-n\alpha} \quad (38)$$

$$=\sum_{k\geq 0}[k\beta][k\gamma]\,\frac{(n\alpha)^k}{k!}\,e^{-n\alpha}=(\nu\alpha\beta)(\nu\alpha\gamma)+\nu\alpha\beta\gamma \quad (39)$$

Par conséquent,

$$\mathrm{Cov.}(R,S)=\nu\alpha\beta\gamma \quad \text{et} \quad \mathrm{Cor.}(R,S)=\sqrt{\beta\gamma} \quad (40)$$

Pour exactement les mêmes raisons que celles pour la situation analogue du paragraphe IV.2.3.2., on a le résultat suivant :

PROPRIETE 4. La matrice des covariances de la suite des indices $\{s(a'_h,a'_j)/1\leq h<j\leq m\}$ aléatoires dans l'h.a.l. N_3 est définie.

IV.2.4. Conjecture relative à la loi de la somme des carrés des indices aléatoires et justification de la réduction globale des similarités

Nous allons considérer ici -pour la comparaison deux à deux d'un ensemble d'attributs descriptifs- la forme la plus simple de l'h.a.l. ; c'est-à-dire N_1. Il est parfaitement équivalent pour la comparaison deux à deux d'un ensemble V de variables numériques de considérer la forme permutationnelle de l'h.a.l. (cf.§IV.2.1.).

$\mathcal{A}'=\{a'_j/1\leqslant j\leqslant m\}$ étant l'ensemble des attributs aléatoires associé à $\mathcal{A}$, nous nous intéressons à la statistique définie par la somme des carrés des indices aléatoires d'association :

$$\{Q^2(a'_j,a'_k)/1\leqslant j<k\leqslant m\}\ , \qquad (41)$$

où Q(a,b) se trouve défini par l'expression (1) du paragraphe III.1.

Il s'agit dans le cadre de l'h.a.l. N_1 d'une somme de m(m-1)/2 carrés de v.a. normales centrées et réduites dont d'ailleurs la matrice des covariances est la matrice unité (cf. § IV.2.2. et § IV.2.3.1.).

Toutefois, la loi jointe n'est pas normale compte tenu de la dépendance évidente entre les différents indices aléatoires d'association, comme on le constatera au paragraphe IV.2.5 suivant où on étudie la situation relative de trois attributs aléatoires. Néanmoins, nous proposons la conjecture suivante:

Conjecture : La distribution asymptotique de probabilité de la somme (41) est une loi du χ^2 à m degrés de liberté.

Nous espérons asseoir prochainement cette conjecture à partir de la simulation de tableaux d'incidence aléatoires dont seront fixées les caractéristiques cardinales des vecteurs lignes qui correspondent à $\mathcal{A}'$ et dont le nombre de colonnes sera égal au nombre n=card(E).

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

PROPRIETE 5. La moyenne et la variance de la statistique (41) sont respectivement égales à μ et à 2μ où $\mu=m(m-1)/2$.

Pour ce qui est de l'espérance mathématique, c'est immédiat puisque sous l'h.a.l. N_1, $Q(a'_j,a'_k)$ est une v.a. (0,1). C'est également clair pour ce qui est de la variance puisque les v.a. $Q(a'_j,a'_k)$ sont deux à deux indépendantes et

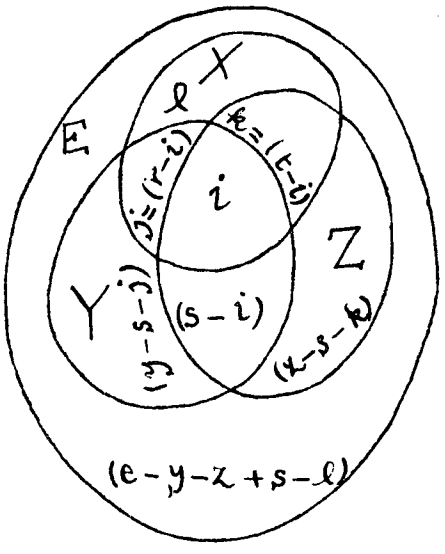
$$\mathcal{V}\left[\sum\{Q^2(a'_j,a'_k)/1\leqslant j<k\leqslant m\}\right]^2=\mu\mathcal{V}\left[Q^4(a'_j,a'_k)\right]+\mu(\mu-1)\mathcal{V}\left[Q^2(a'_j,a'_k)Q^2(a'_{j'},a'_{k'})\right] \quad \text{où}$$

$(j',k')\neq(j,k)$, d'où,

$$=3\mu+\mu(\mu-1) \qquad (42)$$

Revenons alors à la statistique aléatoire de proximité (12). Compte tenu de la très faible liaison entre le numérateur et le dénominateur -dont les carrés ont dans l'h.a.l. N_1, une corrélation de $1/\sqrt{\mu}$ - cette statistique suit une loi de Student généralement assimilable à une loi normale $\mathcal{N}(0,1)$ (μ grand). Ainsi, l'expression (13) se trouve justifiée pour la définition d'une échelle de fréquence mathématique pour la comparaison deux à deux et de façon relative, d'un ensemble de variables descriptives.

IV.2.5. Distribution du triplet $(\text{card}(X\cap Y),\text{card}(X\cap Z),\text{card}(Y\cap Z))$.



Nous allons ici prendre les notations suivantes : e=card(E), x=card(X), y=card(Y) et z=card(Z). Nous allons d'autre part commencer par déterminer cette distribution dans le cadre de l'h.a.l. N_1 ; c'est-à-dire

$\Pr\{|X\cap Y|=r,\ |Y\cap Z|=s,\ |X\cap Z|=t/N_1\}=$

$P_1(r,s,t)$, (43)

pour $\max(0,x+y-e)\leq r\leq\min(x,y)$

$\max(0,y+z-e)\leq s\leq\min(y,z)$

$\max(0,x+z-e)\leq t\leq\min(x,z)$

Nous allons écrire la probabilité (43) sous la forme

$$P_1(r,s,t)=\Pr\{|X\cap Y|=r,\ |X\cap Z|=t/|Y\cap Z|=s\}\ x\ \Pr\{|Y\cap Z|=s\}$$

$$=\Pr\{|X\cap Y|=r,\ |X\cap Z|=t/|Y\cap Z|=s\}x\ \frac{y!z!(e-y)!(e-z)!}{e!s!(y-s)!(z-s)!(e-y-z+s)!} \qquad (44)$$

La probabilité définie par le premier facteur est la même quels que soient les sous-ensembles Y et Z pour lesquels card(Y)=y, card(Z)=z et $\text{card}(Y\cap Z)=s$. Décomposons cette probabilité par rapport aux différentes répartitions de X dont chacune se trouve définie comme suit :

$$\text{card}(X\cap Y\cap Z)=i,\text{card}(X\cap Y\cap Z^c)=j,\text{card}(X\cap Y^c\cap Z)=h \text{ et } \text{card}(X\cap Y^c\cap Z^c)=\ell, \qquad (45)$$

où i,j,k et ℓ sont des entiers positifs ou nuls de somme égale à x. On a

$$\Pr\{|X\cap Y|=r,|X\cap Z|=t/|Y\cap Z|=s\}=\left\{\frac{\binom{s}{i}\binom{y-s}{j}\binom{z-s}{k}\binom{e-y-z+s}{\ell}}{\binom{e}{x}}\Big/(i,j,k,\ ),i,j,k,\ell\geq 0,\ i+j=r,i+k=t,\ell=x-i-j-k.\right\} \qquad (46)$$

Dans ces conditions et après mise en forme, on obtient pour la probabilité (44)

$$P_1(r,s,t)=\sum\left\{\frac{e}{\binom{e}{x}\binom{e}{y}\binom{e}{z}}\Big/\max(0,r+s-y,s+t-z,r+t-x)\leq i\leq\min(r,s,t)\right\}, \qquad (47)$$

où le numérateur est un coefficient multinomial. Nous ne voyons pas pour le moment comment cette loi peut être approchée par une loi multinormale.

Nous allons à présent considérer l'h.a.l. N_2 où x,y et z sont regardés comme les réalisations de trois v.a. binomiales indépendantes de paramètres respectifs $(e,\alpha),(e,\beta)$ et (e,γ), où α,β et γ sont trois proportions. La probabilité $P_1(r,s,t,)$ apparaît alors dans ce cadre comme une probabilité conditionnée, et

$$P_2(r,s,t)=\Pr\{|X\cap Y|=r,\ |Y\cap Z|=s,\ |X\cap Z|=t/N_2\}, \qquad (48)$$

peut se mettre sous la forme, qu'il y a lieu d'intégrer par rapport à x,y et z.

$$\sum\left\{\left(i(r-i)(s-i)(t-i)(y-r-s+i)(z-s-t+i)(x-r-t+i)(e-x-y-z+r+s+t-i)\right)\right.$$
$$\left.\times\ \alpha^{x}\bar{\alpha}^{(e-x)}\beta^{y}\bar{\beta}^{(e-y)}\gamma^{z}\bar{\gamma}^{(e-z)}\Big/\max(0,r+s-y,s+t-z,r+t-x)\leq i\leq\min(r,s,t)\right\},\quad(49)$$

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

Nous allons finalement considérer l'h.a.l. N_3 où e lui-même est regardé comme la réalisation d'une v.a. de Poisson de paramètre n. De la sorte $P_2(r,s,t)$ apparaît comme une probabilité conditionnée par la valeur e de la v.a. de Poisson. Ainsi,

$$P_3(r,s,t)=\Pr\left\{|X\cap Y|=r,\ |Y\cap Z|=s,\ |X\cap Z|=t/N_3\right\}$$
$$=P_2(r,s,t)\times\frac{n^e}{e!}\ \exp.(-n)\qquad(50)$$

Calcul effectué, on peut mettre $P_3(r,s,t)$ sous la forme suivante d'une somme d'une probabilité : produit de lois de Poisson, et qu'il y a lieu de sommer par rapport à x,y,z et e.

$$\sum\left\{\frac{(n\alpha\beta\gamma)^{i}\exp(-n\alpha\beta\gamma)}{i!}\times\frac{(n\alpha\beta\bar{\gamma})^{(r-i)}\exp(-n\alpha\beta\bar{\gamma})}{(r-i)!}\times\frac{(n\bar{\alpha}\beta\gamma)^{(s-i)}\exp(-n\bar{\alpha}\beta\gamma)}{(s-i)!}\right.$$
$$\times\frac{(n\alpha\bar{\beta}\gamma)^{(t-i)}\exp(-n\alpha\bar{\beta}\gamma)}{(t-i)!}\times\frac{(n\bar{\alpha}\bar{\beta}\bar{\gamma})^{(y-s-r+i)}\exp(-n\bar{\alpha}\bar{\beta}\bar{\gamma})}{(y-r-s+i)!}$$
$$\times\frac{(n\bar{\alpha}\bar{\beta}\gamma)^{(z-s-t+i)}\exp(-n\bar{\alpha}\bar{\beta}\gamma)}{(z-s-t+i)!}\times\frac{(n\alpha\bar{\beta}\bar{\gamma})^{(x-r-t+i)}\exp(-n\alpha\bar{\beta}\bar{\gamma})}{(x-r-t+i)!}$$
$$\left.\times\frac{(n\bar{\alpha}\bar{\beta}\bar{\gamma})^{(e-x-y-z+r+s+t-i)}\exp(-n\bar{\alpha}\bar{\beta}\bar{\gamma})}{(e-x-y-z+r+s+t-i)!}\Big/\max(0,r+s-y,s+t-z,r+t-x)\leq i\leq\min(r,s,t)\right\}.$$
(51)

On commence par intégrer par rapport à e pour x,y,z,r,s,t et i fixés, puis par rapport à x,y et z pour r,s,t et i fixés. Finalement, on obtient

$$P_3(r,s,t)=\sum\left\{\frac{(n\alpha\beta\gamma)^{i}\exp(-n\alpha\beta\gamma)}{i!}\times\frac{(n\alpha\beta\bar{\gamma})^{(r-i)}\exp(-n\alpha\beta\bar{\gamma})}{(r-i)!}\right.$$
$$\left.\times\frac{(n\bar{\alpha}\beta\gamma)^{(s-i)}\exp(-n\bar{\alpha}\beta\gamma)}{(s-i)!}\times\frac{(n\alpha\bar{\beta}\gamma)^{(t-i)}\exp(-n\alpha\bar{\beta}\gamma)}{(t-i)!}/0\leq i\leq\min(r,s,t)\right\}.\quad(52)$$

THEOREME. I, J, K et L étant quatre v.a. indépendantes de Poisson de paramètres respectifs $n\alpha\beta\gamma$, $n\alpha\bar{\beta}\gamma$, $n\bar{\alpha}\beta\gamma$ et $n\bar{\alpha}\bar{\beta}\gamma$, la loi du triplet $|X\cap Y|$, $|Y\cap Z|$, $|X\cap Z|$) dans l'h.a.l. N_3 est celle de (I+J, I+K, I+L).

IV.2.6. H.a.l. compatible avec N_2 puis avec N_3 où $\{Q(a'_j,a'_k)/1\leqslant j<k\leqslant m\}$ suit asymptotiquement une loi multinomiale dont la matrice des covariances est définie

IV.2.6. Expression de l'h.a.l. N_{02} et de la suite des v.a. $\{n(a'_j\wedge a'_k)/1\leqslant j<k\leqslant m\}$.

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

$$p(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_j,\ldots,\varepsilon_m)=\prod_{1\leqslant j\leqslant m} p(\varepsilon_j), \quad (53)$$

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

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

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

$$\{n(a'_j\wedge a'_k)/1\leqslant j<k\leqslant 4\} \quad (54)$$

par rapport à celle que nous venons d'introduire

$$\{n(\varepsilon'_1\wedge\varepsilon'_2\wedge\varepsilon'_3\wedge\varepsilon'_4)/(\varepsilon_1,\varepsilon_2,\varepsilon_3,\varepsilon_4)\in\{0,1\}^4\} \quad (55)$$

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

$$\left.\begin{aligned}
X_{12} &= X_{1100}+X_{1101}+X_{1110}+X_{1111}\\
X_{13} &= X_{1010}+X_{1011}+X_{1110}+X_{1111}\\
X_{14} &= X_{1001}+X_{1011}+X_{1101}+X_{1111}\\
X_{23} &= X_{0110}+X_{0111}+X_{1110}+X_{1111}\\
X_{24} &= X_{0101}+X_{0111}+X_{1101}+X_{1111}\\
X_{34} &= X_{0011}+X_{0111}+X_{1011}+X_{1111}
\end{aligned}\right\} \quad (56)$$

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

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

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

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

$$\begin{bmatrix} 0&0&0&0&0&0&0&1&1&1&1\\ 0&0&0&0&0&1&1&0&0&1&1\\ 0&0&0&0&1&0&1&0&1&0&1\\ 0&0&1&1&0&0&0&0&0&1&1\\ 0&1&0&1&0&0&0&0&1&0&1\\ 1&0&0&1&0&0&1&0&0&0&1 \end{bmatrix} \qquad (57)$$

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

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

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

IV.2.6.2. Calcul dans le cadre de N_{02} de la matrice des covariances de $\{n(a'_j,a'_k)/1\leq j<k\leq m\}$.

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

$$\left.\begin{aligned} n(a'\wedge b')&=n(a'\wedge b'\wedge c'\wedge d')+n(a'\wedge b'\wedge c'\wedge \bar{d}')+n(a'\wedge b'\wedge \bar{c}'\wedge d')+n(a'\wedge b'\wedge \bar{c}'\wedge \bar{d}')\\ n(a'\wedge c')&=n(a'\wedge b'\wedge c'\wedge d')+n(a'\wedge \bar{b}'\wedge c'\wedge d')+n(\bar{a}'\wedge \bar{b}'\wedge c'\wedge d')+n(\bar{a}'\wedge \bar{b}'\wedge c'\wedge \bar{d}')\\ n(c'\wedge d')&=n(a'\wedge b'\wedge c'\wedge d')+n(a'\wedge \bar{b}'\wedge c'\wedge d')+n(\bar{a}'\wedge b'\wedge c'\wedge d')+n(\bar{a}'\wedge \bar{b}'\wedge c'\wedge d') \end{aligned}\right\} \qquad (58)$$

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

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

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

$$\text{var.}\left[n(a',b')\right]=n\alpha\beta x\left\{1-\alpha\beta\left[1+2\left(\gamma\bar{\gamma}(1-\delta\bar{\delta})+\delta\bar{\delta}(1-\gamma\bar{\gamma})\right)\right]\right\} \qquad (59)$$

$$\begin{aligned}\text{cov.}\left[n(a',b'),n(a',c')\right]=&n\alpha\beta\gamma\delta\{(1-\alpha\beta\gamma\delta)-\alpha\beta\gamma\bar{\delta}-\alpha\beta\bar{\gamma}\delta-\alpha\bar{\beta}\gamma\delta\}\\&+n\alpha\beta\gamma\bar{\delta}\{-\alpha\beta\gamma\delta+(1-\alpha\beta\gamma\bar{\delta})-\alpha\beta\bar{\gamma}\delta-\alpha\bar{\beta}\gamma\delta-\alpha\bar{\beta}\gamma\bar{\delta}\}\\&+n\alpha\beta\bar{\gamma}\delta\{-\alpha\beta\gamma\delta-\alpha\beta\gamma\bar{\delta}-\alpha\beta\bar{\gamma}\delta-\alpha\bar{\beta}\gamma\delta-\alpha\bar{\beta}\gamma\bar{\delta}\}\\&+n\alpha\beta\bar{\gamma}\bar{\delta}\{-\alpha\beta\gamma\delta-\alpha\beta\gamma\bar{\delta}-\alpha\beta\bar{\gamma}\delta-\alpha\bar{\beta}\gamma\delta-\alpha\bar{\beta}\gamma\bar{\delta}\}\end{aligned}$$

$$\begin{aligned}&=n\alpha\beta\gamma-n\alpha\beta\{\alpha\gamma\}\\&=n\alpha\bar{\alpha}\beta\gamma \qquad (60)\end{aligned}$$

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

$$\begin{aligned}\text{cov.}\left[n(a'\wedge b'),n(c'\wedge d')\right]=&n\alpha\beta\gamma\delta\{1-\alpha\beta\gamma\delta)-\alpha\bar{\beta}\gamma\delta-\bar{\alpha}\beta\gamma\delta-\bar{\alpha}\bar{\beta}\gamma\delta\}\\&+n\alpha\beta\gamma\bar{\delta}\{-\alpha\beta\gamma\delta-\alpha\bar{\beta}\gamma\delta-\bar{\alpha}\beta\gamma\delta-\bar{\alpha}\bar{\beta}\gamma\delta\}\\&+n\alpha\beta\bar{\gamma}\delta\{-\alpha\beta\gamma\delta-\alpha\bar{\beta}\gamma\delta-\bar{\alpha}\beta\gamma\delta-\bar{\alpha}\bar{\beta}\gamma\delta\}\\&+n\alpha\beta\bar{\gamma}\bar{\delta}\{-\alpha\beta\gamma\delta-\alpha\bar{\beta}\gamma\delta-\bar{\alpha}\beta\gamma\delta-\bar{\alpha}\bar{\beta}\gamma\delta\}\end{aligned}$$

$$\begin{aligned}&=n\alpha\beta\gamma\delta-n\alpha\beta\gamma\delta\\&=0 \qquad (61)\end{aligned}$$

On peut remarquer que comme dans le cas de l'h.a.l. N_2 (cf.§ IV.2.3.2.) définie de façon intrinsèque et globale sans référence à une population parente $\mathcal{P}$, on retrouve les mêmes valeurs pour $\text{cov.}\left[n(a'\wedge b'),n(a'\wedge c')\right]$ et pour $\text{cov.}\left[n(a'\wedge b'),n(c'\wedge d')\right]$. Toutefois, $\text{var}\left[n(a'\wedge b')\right]$ change légèrement en diminuant de $n\alpha\beta\left[1-\alpha\beta\right]$ au second membre de (59). De sorte que pour des raisons analogues à celles avancées au paragraphe IV.2.3.2., la matrice des covariances est définie.

THEOREME. Sous l'hypothèse d'absence de liaison N_{02} définie ci-dessus, la distribution asymptotique de la suite des v.a. $\{n(a'_j\wedge a'_k)/1\leqslant j<k\leqslant m\}$ est asymptotiquement normale de matrice des covariances -déterminée par les formules (59),(60) et (61)- définie.

IV.2.6.2. Matrice des covariances dans le cas de l'h.a.l. N_{03} correspondant à un modèle Poissonnien

Dans le cadre d'une telle hypothèse, la loi jointe de $\left\{n(\varepsilon'_1\ \varepsilon'_2\ \ldots\ \varepsilon'_m)/(\varepsilon_1,\ldots,\varepsilon_2,\varepsilon_m)\ \theta\ \{0,1\}^m\right\}$ est un produit de 2^m lois indépendantes de Poisson, où le paramètre de la loi de $n(\varepsilon'_1\ \varepsilon'_2\ \ldots\ \varepsilon'_m)$ est $n\Pi\{\alpha_j^{\varepsilon_j}(\bar{\alpha}_j)^{(1-\varepsilon_j)}/1\leqslant j\leqslant m\}$.

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

$$\begin{aligned}&\text{var.}\left[n(a'\wedge b')\right]=n\alpha\beta\\&\text{cov.}\left[n(a'\wedge b'),n(a'\wedge c')\right]=n\alpha\beta\gamma\delta+n\alpha\beta\gamma\bar{\delta}=n\alpha\beta\gamma \qquad (62)\\&\text{cov.}\left[n(a'\wedge b'),n(c'\wedge d')\right]=n\alpha\beta\gamma\delta\end{aligned}$$

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

$$\begin{array}{c} \\ \alpha\beta \\ \\ \\ \\ \\ \end{array}\begin{bmatrix} 1 & \gamma & \delta & \gamma & \delta & \gamma\delta \\ \beta & 1 & \delta & \beta & \beta\delta & \delta \\ \beta & \gamma & 1 & \beta\gamma & \beta & \gamma \\ \alpha & \alpha & \alpha\beta & 1 & \delta & \delta \\ \alpha & \alpha\gamma & \alpha & \gamma & 1 & \gamma \\ \alpha\beta & \alpha & \alpha & \beta & \beta & 1 \end{bmatrix} \qquad (63)$$

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

$$\left[\begin{array}{c|c|c|c|c|c} 1 & \gamma & \delta & \gamma & \delta & \gamma\delta \\ 0 & 1-\beta\gamma & \delta(1-\beta) & \beta(1-\gamma) & 0 & \delta(1-\beta\gamma) \\ 0 & \gamma(1-\beta) & 1-\beta\delta & 0 & \beta(1-\delta) & \gamma(1-\beta\delta) \\ 0 & \alpha(1-\gamma) & 0 & 1-\alpha\gamma & \delta(1-\alpha) & \delta(1-\alpha\gamma) \\ 0 & 0 & \alpha(1-\delta) & \gamma(1-\alpha) & 1-\alpha\delta & \gamma(1-\alpha\delta) \\ 0 & \alpha(1-\alpha\gamma) & \alpha(1-\delta) & \beta(1-\alpha\gamma) & \beta(1-\alpha\delta) & 1-\alpha\beta\gamma\delta \end{array}\right] \qquad (64)$$

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

THEOREME. Sous l'hypothèse d'absence de liaison N_{03} du modèle Poissonnien, la distribution asymptotique de la suite des v.a.$\{n(a'_j\wedge a'_k)/1\leq j<k\leq m\}$ est asymptotiquement normale de matrice des covariances -déterminée par les formules (62)- définie

IV.2.6.3. Sur deux modes nouveaux de réduction globale des similarités

Ces deux modes sont une conséquence directe de l'analyse des paragraphes précédents (IV.2.6.1., IV.2.6.2. et IV.2.6.3.). Désignons par V_{02}(resp.V_{03}) la matrice des covariances obtenue au paragraphe IV.2.6.2.(resp.IV.2.6.3.) et par $q'={}^t(q(a'_j,a'_k)/1\leq j<k\leq m)$, le vecteur colonne des indices aléatoires "centrés" de proximité $(q(a'_j,a'_k)=[n(a'_j,a'_k)-n\alpha_j\alpha_k])$, on déduit des théorèmes précédents que sous l'hypothèse N_{02}(resp. N_{03}) la v.a. ${}^tq'V_{02}q'$(resp.${}^tq'V_{03}q'$) suit une loi du χ^2 à $\mu=\binom{m}{2}$ degrés de liberté [LANCASTER(1969)].

Si maintenant on désigne par V_{01i} (i=1,2 ou 3) la matrice diagonale des variances $\{var_i[q(a'_j,a'_k)]/1\leq j<k\leq m\}$ où var_i est la variance calculée dans l'h.a.l. N_i, les formules (9) et (10) (§ IV.1.) proposent une réduction globale au moyen de $\{{}^tq\,V_{01i}q/\binom{m}{2}\}^{1/2}$ qu'on justifie au mieux au moyen de la conjecture -pour i=1- du paragraphe IV.2.4. Suite à la précédente analyse, nous pouvons proposer deux autres modes parfaitement justifiés de réduction globale ; le premier au moyen de $\{{}^tqV_{02}q/\binom{m}{2}\}^{1/2}$ et le second, au moyen de $\{{}^tqV_{03}q/\binom{m}{2}\}^{1/2}$.

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

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

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

$$s=\mathrm{card}\left[R(\alpha)\cap R(\beta)\right] \qquad (2)$$

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

$$S=\mathrm{card}\left[R(\alpha')\cap R(\beta')\right] \qquad (3)$$

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

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

Figure 1.

que nous allons illustrer dans deux situations classiques que nous avons déjà évoquées au paragraphe II ci-dessus.

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

De façon plus explicite, notons

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

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

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

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

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

et

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

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

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

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

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

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

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

et où

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

Imaginons à présent que les n_{ij} soient "assez grands", on a

$$\left.\begin{aligned}&s=O_1\left[\frac{n^2}{2}\left(\sum\{p_{ij}^2 / 1\leqslant i\leqslant I,\ 1\leqslant j\leqslant J\}\right)\right]\\ &\lambda=O_1\left[\frac{n}{\sqrt{2}}\Big(\sum_{1\leqslant i\leqslant I} p_{i.}^2\Big)\right],\quad \mu=O_1\left[\frac{n}{\sqrt{2}}\Big(\sum_{1\leqslant j\leqslant J} p_{.j}^2\Big)\right]\\ &\rho=O_1\left[\sqrt{n^3}\Big(\sum_{1\leqslant i\leqslant I} p_{i.}^3\Big)\right],\quad \sigma=O_1\left[\sqrt{n^3}\Big(\sum_{1\leqslant j\leqslant J} p_{.j}^3\Big)\right]\\ &\theta=O_1\left[\frac{n^2}{2}\Big(\Big[\sum_{1\leqslant i\leqslant I} p_{i.}^2\Big]^2\Big)\right],\quad \zeta=O_1\left[\frac{n^2}{2}\Big(\Big[\sum_{1\leqslant j\leqslant J} p_{.j}^2\Big]^2\Big)\right],\end{aligned}\right\} \quad (4)$$

où O_1 indique "se comporte -pour les n_{ij} tendant vers l'infini- comme". Mais, attention, pour le développement de $(\theta\zeta-\lambda^2\mu^2)$, on a exactement

$$O_1(\theta\zeta-\lambda^2\mu^2)=n^3\{\Big[\Big(\sum_i p_{i.}^2\Big)\Big(\sum_j p_{.j}^2\Big)\Big]^2-\Big(\sum_i p_{i.}^2\Big)^2\Big(\sum_j p_{.j}^3\Big)-\Big(\sum_i p_{i.}^3\Big)\Big(\sum_j p_{.j}^2\Big)^2\} \quad (5)$$

Finalement, la partie dominante de la variance se comporte comme

$$n^3\{\Big[\Big(\sum_i p_{i.}^2\Big)^2-\Big(\sum_i p_{i.}^3\Big)\Big]\Big[\Big(\sum_j p_{.j}^2\Big)^2-\Big(\sum_j p_{.j}^3\Big)\Big]\} \quad (6)$$

De sorte, que la forme limite de l'indice d'association $Q(\alpha,\beta)$ est -au facteur (1/2) près-

$$\frac{\sqrt{n}\ \times\ \sum\{(p_{ij}^2-p_{i.}^2\,p_{.j}^2)\,/\,1\leqslant i\leqslant I,\ 1\leqslant j\leqslant J\}}{\sqrt{\Big[\Big(\sum_i p_{i.}^2\Big)^2-\Big(\sum_i p_{i.}^3\Big)\Big]\Big[\Big(\sum_j p_{.j}^2\Big)^2-\Big(\sum_j p_{.j}^3\Big)\Big]}} \quad (7)$$

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

L'étude statistique mentionnée ci-dessus a donc été essentielle pour la découverte de l'expression formelle de l'indice (7).

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

$$T_{\alpha\beta} = \phi^2_{\alpha\beta}/\sqrt{(I-1)(J-1)}, \qquad (8)$$

où

$$\phi^2_{\alpha\beta} = \sum\{(p^2_{ij}/p_{i.}\,p_{.j})\Big/1\leq i\leq I,\ 1\leq j\leq J\}-1, \qquad (9)$$

$$= \chi^2(\alpha,\beta)/n. \qquad (10)$$

L'indice ϕ^2 ou $T_{\alpha\beta}$ est de nature différente de celui se déduisant de (7) ci-dessus (en divisant par $\sqrt{n}$). Néanmoins, on peut considérer que son expression formelle se justifie par une étude statistique préalable au niveau de E, puisqu'on peut montrer que $\chi^2(\alpha',\beta')$ -où (α',β') est un couple de partitions aléatoires indépendantes de types respectifs $t(\alpha)$ et $t(\beta)$- suit une loi du χ^2 à (I-1)(J-1) degrés de liberté.

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

$$R(\alpha)= \sum\{E_i xE_{i'}/1\leq i<i'\leq I\},\ R(\beta)= \sum\{F_j xF_{j'}/1\leq j<j'\leq J\ , \qquad (11)$$

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

$$s=\mathrm{card}[R(\alpha)\cap R(\beta)]= \sum\{n_{ij}n_{i'j'}/1\leq i<i'\leq I,\ 1\leq j<j'\leq J\}. \qquad (12)$$

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

$$\lambda\mu \text{ et } [\lambda\mu+\rho_{cc}\sigma_{cc}+\rho_{ff}\sigma_{ff}+2\rho_{cf}\sigma_{cf}+(\theta\zeta-\lambda^2\mu^2)]. \qquad (13)$$

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

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

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

$$\rho_{ff} = \frac{1}{\sqrt{n(n-1)(n-2)}} \sum \{n_{i.} n^{f}_{i.} (n^{f}_{i.} - 1)/2 \leq i \leq I\} \qquad (14)$$

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

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

où on note

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

On a

$$s = O_1(n^2 \sum \{p_{ij} p_{i'j'} / 1 \leq i < i' \leq I,\ 1 \leq j < j' \leq J\})$$

$$\lambda = O_1(n \sum \{p_{i.} p_{i'.} / 1 \leq i < i' \leq I\})$$

$$\rho_{cc} = O_1(n^{3/2} \sum \{p_{i.} (p^{c}_{i.})^2 / 1 \leq i \leq I\}), \quad \rho_{ff} = O_1(n^{3/2} \sum \{p_{i.} (p^{f}_{i.})^2 / 1 \leq i \leq I\}),$$

$$\rho_{cf} = O_1(n^{3/2} \sum \{p_{i.} p^{c}_{i.} p^{f}_{i.} / 1 \leq i \leq I\}) \text{ et}$$

$$\theta = O_1(n^2 [\sum \{p_{i.} p_{i'.} / 1 \leq i < i' \leq I\}]^2), \qquad (15)$$

où $p^{c}_{i.}$ (resp. $p^{f}_{i.}$) est la proportion correspondante à la fréquence absolue $n^{c}_{i.}$ (resp. $n^{f}_{i.}$).

Il est clair qu'on a des relations analogues pour $\mu, \sigma_{cc}, \sigma_{ff}, \sigma_{cf}$ et ζ, en remplaçant de façon conforme, les $p_{i.}$ par les $p_{.j}$, $1 \leq i \leq I$, $1 \leq j \leq J$.

Toutefois, l'approximation du coefficient $Q(\alpha, \beta)$ (cf. niveau (4) du schéma de la figure 1 ci-dessus), nécessite la détermination de $O_1(\theta\zeta - \lambda^2\mu^2)$. On obtient après calcul,

$$\theta\zeta-\lambda^2\mu^2=O_1(n^3\{4(\sum_{i<i'} p_{i.}p_{i'.})^2(\sum_{j<j'} p_{.j}p_{.j'})^2-(\sum_{i<i'} p_{i.}p_{i'.})^2\Big[\sum_{j<j'} p_{.j}(1-p_{.j})p_{.j'}$$

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

$$-(\sum_{j<j'} p_{.j}p_{.j'})^2\Big[\sum_{i<i'} p_{i.}(1-p_{i.})p_{i'.}+\sum_{i<i'} p_{i.}p_{i'.}(1-p_{i'.})\Big]\}) \quad (16)$$

Ainsi, le numérateur de la forme asymptotique de l'indice $Q(\alpha,\beta)$, peut s'écrire

$$\sqrt{n}\sum\{(p_{ij}p_{i'j'}-p_{i.}p_{i'.}p_{.j}p_{.j'})\Big/1\leqslant i<i'\leqslant I,\ 1\leqslant j<j'\leqslant J\}, \quad (17)$$

et le dénominateur représente la racine carrée de

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

$$+\Big[\sum_i p_{i.}(p^f_{i.})^2\Big]\Big[\sum_j p_{.j}(p^f_{.j})^2\Big]$$

$$+2(\sum_i p_{i.}p^c_{i.}p^f_{i.})(\sum_j p_{.j}p^c_{.j}p^f_{.j})-(\sum_{i<i'} p_{i.}p_{i'.})^2\Big[\sum_{j<j'} p_{.j}(1-p_{.j})p_{.j'}$$

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

$$-(\sum_{j<j'} p_{.j}p_{.j'})^2\Big[\sum_{i<i'} p_{i.}(1-p_{i.})p_{i'.}+\sum_{i<i'} p_{i.}p_{i'.}(1-p_{i'.})\Big] \quad (18)$$

On remarquera que -comme pour la formule (7) ci-dessus de comparaison de deux variables "partition"- l'expression limite de l'indice d'association entre deux variables "préordre total", se présente comme $\sqrt{n}\,\mathcal{G}[p(\alpha\wedge\beta)]$, où $\mathcal{G}[p(\alpha\wedge\beta)]$ est la fonction du tableau $p(\alpha\wedge\beta)$ des proportions $p_{ij}, 1\leqslant i\leqslant I, 1\leqslant j\leqslant J$ (cf.(1) . Cette fonction $\mathcal{G}$ définit parfaitement un indice qui peut être appliqué au tableau des mêmes proportions $\pi(\alpha\wedge\beta)$, mais considéré au niveau de la population totale .

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

Cependant, l'étude développée par Goodman et Kruskal [GOODMAN & KRUSKAL (1963),(1972)] relative à la distribution d'échantillonnage de l'indice calculé au niveau de E -par rapport à sa valeur théorique au niveau de la population - est très intéressante. Nous l'avons reprise (au paragraphe III.2.) pour évaluer la précision de nos indices d'association entre attributs descriptifs. Curieusement, les auteurs précités n'avaient pas considéré le cas de la comparaison d'attributs dont nous avons ci-dessus montré toute la richesse. On peut considérer le même type d'étude pour les indices que nous exprimons ci-dessus ($f[p(\alpha\wedge\beta)]$ et $g[p(\alpha\wedge\beta)]$), mais les calculs deviennent par trop complexes. Une possibilité s'offre toutefois, si on étudie séparément le numérateur et le dénominateur d'un même indice.

Le fait d'introduire le point de vue inférentiel où l'ensemble E (des objets ou individus) est regardé comme la réalisation d'un échantillon aléatoire de taille croissante de $\mathfrak{P}$, nous a permis de nous rendre compte que l'indice de la vraisemblance de la liaison, relativement à une h.a.l. à caractère intrinsèque (au niveau de E), n'acquiert tout son intérêt que lors de la comparaison deux à deux d'un ensemble de variables de même type, où alors il y a lieu au préalable de réduire globalement (cf.§IV), les indices $Q(\alpha,\beta)$ (cf. niveau (4) du schéma de la figure 1 ci-dessus).

La justification de cette réduction globale des similarités que -par nécessité- nous pratiquions depuis longtemps, nous a permis d'aborder des problèmes qui nous paraissent tout à fait originaux et fondamentaux de la statistique non paramétrique multivariée.

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique