MATHÉMATIQUES ET SCIENCES HUMAINES

I. C. LERMAN
R. GRAS
H. ROSTAM

Élaboration et évaluation d'un indice d'implication pour des données binaires. I

Mathématiques et sciences humaines, tome 74 (1981), p. 5-35
<http://www.numdam.org/item?id=MSH_1981__74__5_0>

Math. Sci. hum. (19^{e} année, n°74, *1981*, p.5-35)

ELABORATION ET EVALUATION D'UN INDICE D'IMPLICATION POUR DES DONNEES BINAIRES . I

I.C. LERMAN, R. GRAS et H. ROSTAM(*)

I - INTRODUCTION ; POSITION DU PROBLEME ET PLAN DU TRAVAIL

Le problème de la recherche d'enchaînements successifs entre comportements est un problème général dans les Sciences Humaines (échelle hiérarchique d'attitude en Psychologie, sériation en Archéologie). Le spécialiste cherche en effet à détecter un phénomène qui évolue et à maîtriser la variable qui conditionne cette évolution.

Ce travail est une contribution méthodologique à la solution de ce problème qui s'est présenté ici à nous dans le cadre de la Didactique. Dans ce cadre, on se pose constamment la question : "Dans quelle mesure tel comportement à un stimulus donné a implique tel comportement à un autre stimulus b ?" (pour fixer les idées, a et b peuvent par exemple être définis par deux questions d'un test mathématique et le comportement peut être la réussite).

De ce type de problème, on a surtout cherché à donner une solution globale, rigide et ambitieuse. On élabore à cet effet, un ensemble A de stimuli (ici des questions) par rapport auquel on postule l'existence d'une même variable sous-jacente (ici la complexité). Celle-ci ordonnerait A au sens suivant : $\forall (a,b) \in A \times A, a < b$ si et seulement si un certain comportement (ici la réussite) vis-à-vis de a implique le même comportement vis-à-vis de b. On notera aussi dans ce cas : $a \Rightarrow b$.

Pour tester une telle hypothèse, on cherche relativement à une expérience réelle sur un ensemble E de sujets, à construire l'ordre total sur A

(*) Laboratoire de Statistique, IRISA, Université de Rennes I et IREM de Rennes. "Campus Beaulieu", 35042 RENNES CEDEX.

qui s'"ajuste au mieux" aux comportements observés. On mesure l'adéquation d'un tel ajustement et on confronte l'ordre dégagé avec celui postulé résultant de l'hypothèse d'unidimensionnalité. (cf. [8], [17], [20]).

La réalité en Didactique est trop complexe pour admettre une telle hypothèse simplificatrice. Au lieu d'un ordre total strict sur A, on postulera seulement un préordre total sur A correspondant à une "taxinomie d'objectifs cognitifs" proposée par R. Gras (cf. [7]). Au lieu de chercher à mettre directement en évidence le "meilleur" préordre total à partir du comportement réel d'un ensemble E de sujets, on proposera de bâtir un ordre partiel pondéré, matérialisé par un graphe orienté valué et sans cycles où la valuation portée sur un arc (a,b) de AxA, représentera précisément la mesure du degré d'implication a => b. C'est pour cette raison qu'on appellera ce graphe "graphe d'implication". Un des aspects techniques futurs de ce travail est l'automatisation du dessin de ce graphe où on ne retient que les arcs dont la valuation est supérieure à un seuil donné, seuil défini sur des bases statistiques.

La synthèse de l'information quant à l'analyse de la relation qui existe entre lignes (resp. colonnes) d'un même tableau de données, au moyen d'un graphe valué, est une technique qui existe dans le traitement des données. Mais il s'agit de graphe symétrique (non orienté) où la valuation correspond à un degré de ressemblance et non d'implication (exemple : graphes de similitude de Cl. Flament (cf. [6])). En d'autres termes on analyse une forme symétrique de la relation.

La construction d'un graphe d'implication passe donc nécessairement par l'élaboration d'un indice d'implication. Il y a bien à ce sujet un indice proposé par J. Loevinger mais sa construction trop brute n'est pas sans arbitraire ce qui empêche de comparer sur une même base tous les couples d'items.

S'inspirant de la démarche générale de I.C. Lerman pour la mise au point d'un indice de proximité entre structures finies de même type et plus particulièrement entre variables-attributs (cf. [14] et [15]) , on propose deux indices d'implication qui se réfèrent à une hypothèse statistique d'absence de relation et utilisent une notion de "vraisemblance" de la relation d'implication entre attributs-comportements. Chacun de ces indices permet d'une certaine façon, d'évaluer par rapport à une échelle de probabilité établie dans l'hypothèse d'absence de liaison, l'intensité d'une implication

donnée. On a ainsi une base commune et uniforme d'évaluation d'un même type d'implication associé à un même indice.

Chacun de ces indices sera, comme l'indice de proximité de I.C. Lerman entre attributs, un indice de comparaison entre parties d'un même ensemble. Pour cette raison, nous commencerons par reprendre de façon précise et plus complète que nous l'avions encore fait, le principe de l'élaboration de cet indice de proximité dans le cas de la comparaison entre parties et les différentes formes de cet indice qui correspondent en fait à différentes formes de l'hypothèse d'absence de lien de référence (cf. § II).

Au paragraphe III nous présenterons les deux indices et nous chercherons à les analyser de façon comparative.

Cet article est donc consacré à l'élaboration d'un indice d'implication; il sera directement suivi au prochain numéro d'un article où on étudie le problème de l'évaluation du graphe d'implication associé à un même indice. Pour avoir une vue d'ensemble du travail, introduisons déjà le contenu du prochain article.

Nous commencerons (§ IV, 2) par donner le dessin des deux graphes d'implication respectivement associés à chacun des deux indices dans le cadre d'un exemple réel fourni par un test mathématique, autour du concept de la symétrie centrale, formé d'un ensemble A de 40 questions et subi par un ensemble E de 401 élèves de la classe de 4ème, en 1978. Des chaînes d'un même graphe, le didacticien donne une interprétation spécifique par rapport à la taxinomie d'objectifs cognitifs et à travers une analyse des tâches.

La relation entre le préordre total sur A défini par la taxinomie et le graphe d'implication (relation d'ordre pondérée sur A) est évaluée à partir d'un indice très général de comparaison entre deux relations pondérées sur un même ensemble, indice se référant également à une hypothèse d'absence de lien (cf. [12], [14] et [21]) (§ V). Nous rappellerons l'expression de ce dernier indice qui peut jouer le rôle d'un critère permettant, au moyen d'un algorith-

me d'échanges, de substituer au préordre total défini par la taxinomie, un préordre total qui s'ajuste "au mieux" au graphe d'implication (paragraphe VI).

II - LES TROIS FORMES FONDAMENTALES DE L'HYPOTHESE D'ABSENCE DE LIEN (h.a.l.) ; INDICES DE PROXIMITE ASSOCIES

II.1 - Indice "brut" de proximité et h.a.l.

Il s'agit ici pour nous de rappeler l'élaboration de notre indice de proximité dans le cas particulier de la comparaison de parties d'un même ensemble fini, qu'on notera E. Pour fixer les idées,E définira l'ensemble des individus.

Relativement à un couple de parties (E(a),E(b)) représentant en l'occurrence un couple (a,b) d'attributs de description où E(a) (resp. E(b)) est le sous-ensemble des sujets possédant l'attribut a (resp. b), on introduit le cardinal

$$s = n(a \wedge b) = \mathrm{card}(E(a) \cap E(b)) \qquad (1)$$

qui définit le nombre d'individus possédant chacun des deux attributs a et b. Nous appelons s "indice brut de proximité".

s doit jouer un rôle important dans la construction de l'indice de proximité définitif entre les deux attributs a et b (c'est-à-dire, entre E(a) et E(b)) ; en effet, la présence commune de a et b chez un même individu est une indication positive quant à leur association ou leur ressemblance. Mais seule, la valeur de s est un indicateur certainement biaisé de la similarité entre les deux attributs : il suffit en effet que a et b soient fréquents (resp. rares) pour trouver une valeur de s relativement grande (resp. petite) et ceci indépendamment de la position relative de E(a) et E(b).

D'où l'idée de situer la valeur de l'indice brut s par rapport à la valeur "attendue" du cardinal de l'intersection entre deux parties aléatoires X et Y où X (resp. Y) est associée à E(a) (resp. E(b)), selon un modèle probabiliste à caractère uniforme et respectant d'une "certaine façon" la caractéristique cardinale de E(a) (resp. E(b)). Qui dit "valeur attendue" dit "distribution" ; il s'agit en fait de situer s par rapport à la loi de la variable aléatoire (v.a.) S = card (X ∩ Y) et l'indice que nous proposons prend

la forme générale

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

où N est l'hypothèse d'absence de lien qui définit l'association :

$$(E(a),E(b)) \longrightarrow (X,Y) \qquad (3)$$

En d'autres termes , les deux attributs sont jugés d'autant plus voisins que le nombre d'individus les possédant également est invraisemblablement grand par rapport à l'hypothèse N d'absence de liaison. On introduit donc une notion de vraisemblance dans celle de ressemblance. L'indice P(a,b) se réfère à une échelle de probabilité et est compris entre 0 et 1.

Compte tenu de la référence à la loi normale qui est en général une très bonne approximation de la loi de S, le calcul de l'indice (2) passe par l'expression d'un indice "centré réduit" par rapport à l'hypothèse d'absence de lien N ; soit

$$q(a,b) = (s - \mathcal{E}(S))/\sigma(S) \qquad (4)$$

où $\mathcal{E}(S)$ et $\sigma(S)$ sont respectivement la moyenne et l'écart-type de S dans le cadre de N.

Les différentes formes de l'hypothèse d'absence de lien se distinguent dans leurs manières de respecter les caractéristiques cardinales : n(a) = card(E(a)), n(b) = card(E(b)) et n = card(E). Pour chacune d'entre elles, l'indice centré $(s - \mathcal{E}(S))$ reste le même mais la variance $\sigma^2(S)$ diffère d'un cas à l'autre.

On peut dans ces conditions croire que les analyses des données (par exemple : classification hiérarchique ou analyse factorielle) conformément à l'un ou à l'autre de ces indices, sont quasiment équivalentes. Il n'est est rien; sans donner des résultats qui se contredisent, c'est une forme plutôt qu'une autre de l'h.a.l. qui fournit la synthèse la plus raffinée et la plus cohérente dans les nuances qu'elle fait apparaître (cf.[13]). Nous nous permettons d'autant plus d'insister sur ce point que le mathématicien a trop tendance à croire que dès que deux métriques sont équivalentes (au sens topologique du terme), elles doivent fournir des analyses de données équivalentes.

Nous exprimerons chacune des trois h.a.l. en termes de choix d'un élément aléatoire dans l'ensemble des parties d'un ensemble.

II.2 - Modèle aléatoire 1 de choix (h.a.1. N_1) et indice associé

Soient E un ensemble fini de cardinal n et D une partie de E de cardinal d. Par définition, l'h.a.1. N_1 associe à D un élément aléatoire U dans l'ensemble, muni d'une probabilité uniformément répartie, des parties de E de même cardinal, celui d = card(D). En d'autres termes, en considérant le simplexe 2^E des parties de E, le modèle affecte toute la probabilité, en la répartissant uniformément, sur un même niveau, celui défini par l'ensemble des parties de cardinal d. Dans ces conditions, pour ce modèle

$$\Pr\{U = U_0 / N_1\} = \begin{cases} 0 \text{ si } \mathrm{card}(U_0) \neq d \\ 1/\binom{n}{d} \text{ si } \mathrm{card}(U_0) = d \end{cases} \qquad (5)$$

où $\binom{n}{d}$ est le coefficient binomal bien connu.

Relativement au couple (E(a),E(b))de parties représentant un couple (a,b) d'attributs, l'h.a.1. N_1 peut avoir une forme unilatérale ; par exemple en fixant E(a) et en associant à E(b) une partie aléatoire Y, conformément au modèle. Ou bien, en fixant E(b) et en associant à E(a) une partie aléatoire X. X(resp. Y) est un élément aléatoire dans l'ensemble, muni d'une probabilité uniforme, des parties de E de même cardinal n(b) (resp. n(a)).

On associe ainsi à l'indice brut $s = n(a \wedge b)$ (cf. (1)) les deux v.a. duales

$$S(a) = \mathrm{card}(E(a) \cap Y) \text{ et } S(b) = \mathrm{card}(X \cap E(b)) \qquad (6)$$

On n'a guère à se préoccuper du choix de celle parmi ces deux v.a. qui doit jouer le rôle de la v.a.S de la formule (2) ci-dessus ; en effet :

Propriété 1 : Les distributions de S(a) et de S(b) sont identiques.

On a

$$\Pr\{S(a) = k/N_1\} = \frac{\binom{n(a)}{k}\binom{n(\bar{a})}{n(b)-k}}{\binom{n}{n(b)}} = \frac{\binom{n(b)}{k}\binom{n(\bar{b})}{n(a)-k}}{\binom{n}{n(a)}} = \Pr\{S(b) = k/N_1\} \qquad (7)$$

où $0 \leqslant k \leqslant n(a) \wedge n(b)$.

On a dans cette formule noté $n(\bar{a}) = n-n(a)$ et $n(\bar{b}) = n-n(b)$.

La loi de probabilité commune est hypergéométrique de moyenne et de variance :

$$\mu_1 = \frac{n(a)n(b)}{n} \quad \text{et} \quad \sigma_1^2 = \frac{n(a)n(\bar{a})n(b)n(\bar{b})}{n^2(n-1)} \tag{8}$$

L'indice centré réduit $q_1(a,b) = (s-\mu_1/\sigma_1)$ peut se mettre ici sous la forme

$$q_1(a,b) = \sqrt{n-1} \times \frac{[n(a\wedge b)n(\bar{a}\wedge\bar{b}) - n(a\wedge\bar{b})n(\bar{a}\wedge b)]}{\sqrt{n(a)n(\bar{a})n(b)n(\bar{b})}} \tag{9}$$

où les cardinaux $n(a\wedge b) = \text{card}(E(a) \cap E(b))$, $n(a\wedge\bar{b}) = \text{card}(E(a) \cap E(\bar{b}))$, $n(\bar{a}\wedge b) = \text{card}(E(\bar{a}) \cap E(b))$ et $n(\bar{a}\wedge\bar{b}) = \text{card}(E(\bar{a}) \cap E(\bar{b}))$ sont ceux des classes du croisement des deux partitions de E, chacune en deux classes : $\{E(a),E(\bar{a})\}$ et $\{E(b),E(\bar{b})\}$, conformément au tableau de contingence 2×2 suivant :

		1 : $E(b)$	0 : $E(\bar{b})$	
1	$E(a)$	$n(a\wedge b)$	$n(a\wedge\bar{b})$	$n(a)$
0	$E(\bar{a})$	$n(\bar{a}\wedge b)$	$n(\bar{a}\wedge\bar{b})$	$n(\bar{a})$
		$n(b)$	$n(\bar{b})$	

(10)

L'indice $q_1(a,b)$ n'est autre que le coefficient d'association de K. PEARSON dont le carré est la statistique du chi-deux (χ^2) associé au tableau de contingence précédent. Cette dernière statistique, rappelons-le, se met sous la forme

$$\chi^2 = \sum \left\{ \left[\frac{n(i,j) - (n(i,.)n(.,j)/n}{\sqrt{n(i,.)n(./j)/n}}\right]^2 / (i,j) \in \{0,1\}^2 \right\} \tag{11}$$

où i=1 (resp. 0) code a (resp. $\bar{a}$) et j=1 (resp. 0) code b (resp. $\bar{b}$). Ainsi

$$\frac{n(i,j) - (n(i,.)n(.,j)/n)}{\sqrt{n(i,.)n(.,j)/n}} \tag{12}$$

définit la contribution "orientée" de la case (i,j) à la statistique du χ^2, $0 \leqslant i,j \leqslant 1$.

On peut vérifier que ces contributions orientées peuvent se mettre sous la forme suivante :

$$\begin{cases} \text{case}(1,1) : [n(a\wedge b)n(\bar{a}\wedge\bar{b}) - n(a\wedge\bar{b})n(\bar{a}\wedge b)]/[nn(a)n(b)]^{1/2} \\ \text{case}(1,0) : [n(a\wedge\bar{b})n(\bar{a}\wedge b) - n(a\wedge b)n(\bar{a}\wedge\bar{b})]/[nn(a)n(\bar{b})]^{1/2} \\ \text{case}(0,1) : [n(\bar{a}\wedge b)n(a\wedge\bar{b}) - n(\bar{a}\wedge\bar{b})n(a\wedge b)]/[nn(\bar{a})n(b)]^{1/2} \\ \text{case}(0,0) : [n(\bar{a}\wedge\bar{b})n(a\wedge b) - n(\bar{a}\wedge b)n(a\wedge\bar{b})]/[nn(\bar{a})n(\bar{b})]^{1/2} \end{cases}$$

Compte tenu de la forme de l'expression (9), on peut voir qu'on aurait abouti exactement au même indice centré réduit $q_1(a,b)$ si, au lieu de partir de l'indice brut $s = n(a\wedge b)$, dit "nombre d'associations positives" et contenu de la case(1,1) du tableau (10), on partait de l'indice brut $t = n(\bar{a}\wedge\bar{b})$, "nombre d'associations négatives" et contenu de la case(0,0) du tableau (10). D'autre part, on aurait abouti à un indice exactement opposé si on avait démarré d'un indice brut $u = n(a\wedge\bar{b})$ (contenu de la case(1,0) du tableau (10)).

D'ailleurs, en désignant par $\upsilon(i,j)$ le numérateur de la contribution orientée de la case (i,j) à la statistique du χ^2, $0 \leq i,j \leq 1$, (cf. formules (13)), on a

$$\upsilon(1,1) = \upsilon(0,0) = -\upsilon(1,0) = -\upsilon(0,1) \qquad (14)$$

Le modèle aléatoire 1 (h.a.1.N_1) sera dit "hypergéométrique".

II.3 - Modèle aléatoire 2 de choix (h.a.1.N_2) et indice associé

Comme ci-dessus, considérons une partie fixée D de cardinal d d'un ensemble fini E de cardinal n. Pour faire choix aléatoire d'une partie U associée à D, nous munissons l'ensemble $\mathcal{P}(E)$ des parties de E d'une mesure de probabilité. Alors que dans le modèle aléatoire 1, la mesure de probabilité était concentrée sur un seul niveau du simplexe $\mathcal{P}(E)$, elle sera ici répartie de façon plus diffuse sur les différents niveaux, et le modèle aléatoire 2 (h.a.1.N_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, considérons la v.a. K indice d'un même niveau et cardinal commun de toutes les parties de ce niveau de $\mathcal{P}(E)$. On pose

$$\Pr\{K = k/N_2\} = \binom{n}{k}\delta^k(1-\delta)^{n-k} \qquad (15)$$

où δ est la proportion d/n ; $0 \leq k \leq n$.

. Pour le choix aléatoire d'un élément d'un même niveau k, la probabilité (5) affectée à ce niveau est uniformément répartie sur l'ensemble des $\binom{n}{k}$ points de ce niveau (dont chacun représente une partie du cardinal k) ; chaque point sera de la sorte chargé de la probabilité $\delta^k(1-\delta)^{n-k}$.

Ainsi :

$$\Pr\{U = U_o/N_2,\ c = \mathrm{card}(U_o)\} = \delta^c(1-\delta)^{n-c} \qquad (16)$$

Relativement au couple (E(a),E(b)) de parties représentant un couple (a,b) d'attributs, l'h.a.l.N_2 a ici nécessairement une forme symétrique et globale où directement on associe un couple (X,Y) de parties aléatoires indépendantes. X (resp. Y) est un élément aléatoire dans l'ensemble $\mathcal{P}(E)$ des parties de E, muni de la mesure de probabilité associée au modèle N_2 où δ est remplacé par p(a) = n(a)/n (resp. par p(b) = n(b)/n). Ainsi l'espace de référence de (X,Y) est $\mathcal{P}(E)\times\mathcal{P}(E)$, mais la mesure de probabilité dont on munit le premier ensemble facteur est différente de la mesure de probabilité dont on munit le second ensemble de facteur.

Propriété 2 : La loi de probabilité de la v.a. $\mathrm{card}(X \cap Y)$ est binomiale de paramètre $(n, \pi = p(a)p(b))$.

Preuve :

On a, compte tenu du modèle N_2,

$$\Pr\{\mathrm{card}(X) = k,\ \mathrm{card}(Y) = h\} = \binom{n}{k}p(a)^k p(\bar{a})^{n-k}\binom{n}{h}p(b)^h p(\bar{b})^{n-h} \qquad (17)$$

où on a noté $p(\bar{a}) = 1-p(a)$ et $p(\bar{b}) = 1-p(b)$. D'autre part,

$$\Pr\{\mathrm{card}(X \cap Y) = s/\mathrm{card}(X) = k,\ \mathrm{card}(Y) = h\} = \frac{\binom{k}{s}\binom{n-k}{h-s}}{\binom{n}{h}} \qquad (18)$$

Il s'agit en effet d'une probabilité hypergéométrique déjà définie dans le cadre du modèle N_1 ; elle n'a de sens ici que pour

$$h-s \leq n-k\ ;$$

c'est-à-dire :

$$h+k-s \leq n. \qquad (19)$$

La probabilité cherchée se met donc sous la forme

$$\sum_{G} \binom{k}{s}\binom{n-k}{h-s}\binom{n}{k} p(a)^{k} p(b)^{h} p(\bar{a})^{n-k} p(\bar{b})^{n-h}$$

où G est l'ensemble de sommation : (20)

$$G = \{(k,h)/k \geq s,\ h \geq s \text{ et } h+k-s \leq n\}$$

En faisant le changement de variables

$u = k-s$ et $v = h-s$, G sera défini par

$$\{(u,v)/u \geq 0,\ v \geq 0 \text{ et } u+v \leq (n-s)\}$$

et la probabilité cherchée se met sous la forme

$$\sum_{0\leq u\leq (n-s),\ 0\leq v\leq (n-s)-u} \ \sum \frac{n!}{s!u!v!(n-u-v-s)!}\, p(a)^{s+u} p(b)^{s+v} p(\bar{a})^{(n-s-u)} p(\bar{b})^{(n-s-v)} \qquad (21)$$

Or, on a

$$\sum_{0\leq v\leq (n-s-u)} \frac{(n-s-u)!}{v!(n-s-u-v)!}\, p(b)^{v} p(\bar{b})^{(n-s-u-v)} = 1$$

Il en résulte la simplification suivante de l'expression (21)

$$\sum_{o\leq u\leq (n-s)} \frac{n!}{s!u!(n-s-u)!}\, p(a)^{s+u}\, p(b)^{s}\, p(\bar{a})^{(n-s-u)} p(\bar{b})^{u}$$

$$= \Big(\sum_{o\leq u\leq (n-s)} \frac{(n-s)!}{u!(n-s-u)!} [p(a)p(\bar{b})]^{u} p(\bar{a})^{(n-s-u)}\Big) \times \Big(\frac{n!}{(n-s)!s!} [p(a)p(b)]^{s}\Big)$$

Or, le premier facteur se met sous la forme

$$[p(a)p(\bar{b}) + p(\bar{a})]^{(n-s)} = [1 - p(a)p(b)]^{(n-s)} \qquad (22)$$

D'où le résultat annoncé.

On comprendra aisément dans ces conditions pourquoi nous appelons ce modèle aléatoire de choix N_2, "modèle binomial". Ce dernier aurait certes pu être présenté de façon plus élémentaire au moyen d'un tirage aléatoire avec remise dans l'urne définie par l'ensemble E des objets ou individus où on extrait élément par élément et où pour chacun, la probabilité de faire partie de X (resp. Y) est p(a) (resp.p(b)) , l'appartenance à X étant indépendante de celle à Y. Nous avons préféré donner cette présentation qui, d'une part, donne une vision plus globale du modèle et qui, d'autre part, permet une approche uniforme pour les différentes formes de l'indice de ressemblance ou de proximité.

La moyenne et la variance de la v.a. binomiale card(X∩Y) sont respectivement :

$$\mu_2 = np(a)p(b) = \frac{n(a)\ n(b)}{n} \quad \text{et} \quad \sigma_2^2 = \frac{n(a)n(b)}{n}\ [1 - p(a)p(b)] \qquad (23)$$

D'où l'expression de l'indice centré réduit·

$$q_2(a,b) = \frac{n(a \wedge b) - (n(a)n(b)/n)}{\sqrt{\left[\frac{n(a)n(b)}{n}\right]\left[1 - p(a)p(b)\right]}} \qquad (24)$$

dont le numérateur est le même que celui de $q_1(a,b)$ puisque $\mu_1 = \mu_2$; d'autre part,

$$\sigma_2^2 = \sigma_1^2 \times \frac{p(\bar{a})p(\bar{b})}{[1 - p(a)p(b)]} \qquad (25)$$

(cf. formule (8)).

II.4 - Modèle aléatoire 3 de choix (h.a.l.N_3) et indice associé

D désigne toujours un sous-ensemble donné de cardinal d d'un ensemble fini E de cardinal n. Le modèle précédent permettant de définir la partie aléatoire U associée à D pouvait être exprimé en deux pas. Nous aurons besoin ici de trois pas pour préciser le modèle de choix aléatoire.

Contrairement aux deux modèles précédents, on regardera ici E comme la réalisation d'un ensemble aléatoire $\mathcal{E}$ dont le cardinal est une variable aléatoire N de Poisson de paramètre n :

$$\Pr\{\mathcal{N} = m\} = \frac{n^m}{m!} e^{-n} \qquad (26)$$

pour tout m de l'ensemble $\mathbb{N}$ des entiers.

Conditionnellement à $\mathcal{E} = F$, les deux pas suivants du modèle sont définis comme pour le modèle 2 ci-dessus. Par conséquent

$$\Pr\{K = \mathrm{card}(U)=k/_{N_3}, \mathcal{E}=F\} = \binom{m}{k} \eta^k (1-\eta)^{m-k} \qquad (27)$$

si m = card(F) est supérieur à k , et=0 sinon. Dans cette formule $\eta = d/m$.

De plus,

$$\Pr\{U = U_o/_{N_3}, \mathcal{E}=F, c=\mathrm{card}(U_o)\} = \eta^k (1-\eta)^{m-k} \qquad (28)$$

Ici, c'est au tri-uple d'ensembles (E , E(a),E(b)) où E(a) et E(b) sont deux parties de E, qu'on associe un tri-uple ($\mathcal{E}$,X,Y) d'ensembles aléatoires où X et Y sont deux parties aléatoires indépendantes d'un ensemble aléatoire $\mathcal{E}$ dont on se contente de spécifier la loi de probabilité du cardinal $\mathcal{N}$.

Conformément à ce qui précède, on a

$$\Pr\{\mathrm{card}(X \cap Y) = s/_{N_3}, \mathcal{N}=m\} = \binom{m}{s} \pi^s (1-\pi)^{m-s} \qquad (29)$$

où $\pi = p(a)p(b)$ avec $p(a)=n(a)/m$ et $p(b) = n(b)/m$.

Cette probabilité n'est définie que pour $n(a) \vee n(b) \leq m$ et $s \leq n(a) \wedge n(b)$; on la pose égale à 0 autrement.

Dans ces conditions (compte tenu de (26)) on a

$$\Pr\{\mathrm{card}(X \cap Y) = s/_{N_3}\} = \sum_{m \geq s} \binom{m}{s} \pi^s (1-\pi)^{m-s} \frac{n^m}{m!} e^{-n} \qquad (30)$$

$$= \frac{(n\pi)^s}{s!} e^{-n\pi} \left(\sum_{m \geq s} \frac{[n(1-\pi)]^{m-s}}{(m-s)!} e^{-n(1-\pi)} \right)$$

La somme à l'intérieur de la parenthèse est une somme totale des probabilités d'une loi de Poisson et vaut par conséquent 1.

D'où le résultat

Propriété 3 - La loi de probabilité de la v.a. card($X \cap Y$) dans le cadre du modèle N_3 est de Poisson de paramètre $n\pi$ où $\pi = p(a)p(b)$.

Dans le cadre de ce modèle aléatoire de choix qui sera appelé "modèle Poissonien", l'indice centré réduit se met sous la forme

$$q_3(a,b) = \frac{n(a \wedge b) - (n(a)n(b)/n)}{\sqrt{n(a)n(b)/n}} \qquad (31)$$

Le numérateur de l'indice q_3 est le même que celui de q_1 (resp. q_2) ; d'autre part, pour la variance σ_3^2, on a

$$\sigma_3^2 = \sigma_1^2 \; p(\bar{a})p(\bar{b}) = \sigma_2^2 \; [1 - p(a)p(b)] \qquad (32)$$

On peut remarquer que l'indice $q_3(a,b)$ n'est autre que la contribution orientée de la case (1,1) à la statistique du χ^2 attachée au tableau (10) précédent.

II.5 - Comparaison des trois indices

Pour rappeler les modèles par rapport auxquels ils ont été établis, nous appellerons ici respectivement q_h, q_b et q_p les indices notés précédemment q_1, q_2 et q_3 (h pour "hypergéométrique", b pour "binomial" et p pour "poisson"). Ces trois indices peuvent, au coefficient $\sqrt{n}$ près, se mettre sous la forme

$$\begin{cases} q_h(a,b) = [p(a \wedge b)p(\bar{a} \wedge \bar{b}) - p(a \wedge \bar{b})p(\bar{a} \wedge b)] / [p(a)p(b)p(\bar{a})p(\bar{b})]^{1/2} \\ q_b(a,b) = [p(a \wedge b)p(\bar{a} \wedge \bar{b}) - p(a \wedge \bar{b})p(\bar{a} \wedge b)] / [p(a)p(b)\{1-p(a)p(b)\}]^{1/2} \\ q_p(a,b) = [p(a \wedge b)p(\bar{a} \wedge \bar{b}) - p(a \wedge \bar{b})p(\bar{a} \wedge b)] / [p(a)p(b)]^{1/2} \end{cases}$$

où $p(a \wedge b) = n(a \wedge b)/n$, $p(\bar{a} \wedge \bar{b}) = n(\bar{a} \wedge \bar{b})/n$,

$p(a \wedge \bar{b}) = n(a \wedge \bar{b})/n$, $p(\bar{a} \wedge b) = n(\bar{a} \wedge b)/n$.

L'indice q_b est en quelque sorte "intermédiaire" entre les indices q_h et q_p ; en effet, on a

$$q_p(a,b) < q_b(a,b) < q_h(a,b) \qquad (34)$$

pour tout couple d'attributs (a,b) de AxA.

Pour le voir, il suffira de considérer les dénominateurs respectifs des seconds membres des formules (33) où on a

$$1 > 1 - p(a)p(b) > p(\bar{a})p(\bar{b}) = [1 - p(a)][1 - p(b)] \qquad (35)$$

pour $0 < p(a)p(b) < 1$.

Il y a seulement lieu de comparer les deux indices extrêmes q_p et q_h qui donnent deux points de vue suffisamment différents pour la conception de l'indice de proximité

Propriété 4 - Si $p(a) + p(b) < 1$, on a

$$q_p(a,b) > q_p(\bar{a},\bar{b}) \; ; \qquad (36)$$

alors que

$$q_h(a,b) = q_h(\bar{a},\bar{b}). \qquad (37)$$

C'est immédiat parce que

$$p(a) + p(b) < 1 \iff p(a)p(b) < p(\bar{a})p(\bar{b})$$

La propriété (37) de symétrie correspond à la remarque déjà faite au paragraphe II.2 où nous avons signalé que dans le cadre du modèle N_1 on aboutit au même indice centré réduit si, au lieu de partir de l'indice brut formé du nombre $s = n(a \wedge b)$ d' "associations positives", on partait du nombre $t = n(\bar{a} \wedge \bar{b})$ d'"associations négatives".

Le mathématicien épris de symétrie aura peut-être tendance à préférer l'indice q_h à celui q_p. En fait, cette symétrie est un défaut pour l'approche de la "mesure" de la perception de la ressemblance entre attributs. En effet, si les conditions sont "égales par ailleurs", l'association entre attributs rares doit être plus ponctuée que celle entre attributs fréquents. C'est précisément ce que réalise le modèle N_3 de l'h.a.l. où, en considérant les indices seulement centrés :

$$c_p(a,b) = n(a \wedge b) - [n(a)n(b)/n]$$

et (38)

$$c_p(\bar{a},\bar{b}) = n(\bar{a} \wedge \bar{b}) - [n(\bar{a})n(\bar{b})/n]$$

qui sont égaux ; les v.a. respectivement associées $C_p(a,b)$ et $C_p(\bar{a},\bar{b})$ de même moyenne nulle, sont telles que

$$\text{var. } C_p(a,b) = \frac{p(\bar{a})\ p(\bar{b})}{p(a)\ p(b)} \text{ var } C_p(\bar{a},\bar{b})$$

D'où

$$\text{var. } C_p(a,b) < \text{var. } C_p(\bar{a},\bar{b}) \qquad (39)$$

pour $p(a)+p(b)< 1$.

Ainsi, pour $p(a) + p(b)< 1$, au delà d'un certain seuil, le degré d'invraisemblance de la grandeur relative d'une valeur donnée de l'indice centré est sensiblement plus notable dans le cas où cette valeur concerne $C_p(a,b)$ par rapport au cas où elle concerne $C_p(\bar{a},\bar{b})$.

Il est d'autant plus naturel de préférer l'indice q_p à celui q_h que de la sorte, on se rapproche du point de vue de la théorie de l'Information où c'est l'évènement rare (ici, avoir des associations positives sur le couple d'attributs (a,b)) qui apporte davantage d'information que celui plus fréquent (ici, avoir des associations positives sur le couple d'attributs $(\bar{a},\bar{b})$).

D'autre part, dans la pratique, le spécialiste des données qui élabore l'ensemble A de ses attributs descriptifs à partir d'un ensemble de caractères de description à deux modalités chacun, retient de chacun des caractères l'attribut défini par la modalité la plus rare, généralement considérée comme la plus significative. Relativement à un couple (a,b) d'attributs orientés ainsi obtenus, c'est plus la présence commune que l'absence commune chez un même sujet qui est significative de l'association entre les deux attributs. Il est dans ces conditions naturel de demander à l'indice la propriété définie par l'inégalité (36). Le dénominateur de l'indice q_p permet, de façon plus nette que ne le fait le dénominateur de q_h, de ponctuer l'effet de la rareté commune de a et de b , la fonction $f(p) = p$ tendant plus rapidement vers zéro que ne le fait la fonction $g(p) = p(1 - p)$, pour p positif tendant vers zéro.

Nous avons enfin comparé les résultats obtenus par chacun des deux indices q_h et q_p dans le cadre de la classification hiérarchique d'un ensemble A d'attributs par l'algorithme de la vraisemblance des liens (A.V.L.)(cf.[9]). Les résultats sont pleinement comparables avec parfois des accents différents. Toutefois, c'est plutôt q_p qui donne les résultats les plus nuancés et les plus cohérents dans leurs nuances respectives. Signalons ici à cet égard que cet algorithme prend en charge les indices sous la forme (2) se référant à une échelle de probabilité. Compte tenu de la variance importante des unités de données en Sciences Humaines, après le calcul de l'indice q(a,b) centré et réduit "localement" (cf. formule (4)) mais avant référence à une échelle de probabilité faisant appel à la fonction de répartition de la loi normale centrée réduite, on procède à un centrage et réduction globale des similarités , remplaçant les q(a,b) par les $q'(a,b)=[q(a,b)-\bar{q}]/\sigma_q$ où q et σ_q^2 sont la moyenne et la variance de la distribution

$$\{q(a,b)/\{a,b\} \in P_2(A)\} \quad ; \qquad (40)$$

$P_2(A)$ étant l'ensemble des parties à deux éléments de A. Cette opération qui nous rapproche au niveau global de la distribution des similarités sur l'ensemble des paires de l'h.a.l., permet de mieux mettre en évidence les différences entre les valeurs des indices en utilisant la partie la plus discriminante de la loi normale.

Signalons enfin que, dans notre extension de l'indice à la comparaison des variables d'un autre type , c'est surtout une forme de même nature que celle "hypergéométrique" qui a été jusqu'à présent mise en oeuvre (cf.[15]), tout en modulant l'expression retenue de la variance de l'indice.

III - INDICE D'IMPLICATION

III.1 - L'Indice de J. Loevinger ; introduction aux nouveaux indices

Relativement à deux comportements a et b, définis par exemple à partir de deux questions d'un test mathématique, l'hypothèse "la réussite a implique la réussite b", notée a => b, se trouve parfaitement vérifiée au niveau d'un échantillon formant un ensemble E de sujets si et seulement si

$$E(a) \subset E(b) \qquad (1)$$

où, rappelons-le, E(a) (resp. E(b)) est l'ensemble des sujets ayant réussi l'item a (resp. b).

Dans le cas où a et b représentent deux questions d'un test mathématique, la condition (1) se trouve réalisée de façon suffisante dans deux cas :

(i) le processus mental nécessaire à la solution de b est implicite pour résoudre a ; en d'autres termes, c'est un des composants de la solution de a.

(ii) le degré de complexité de la tâche a est sensiblement plus grand que celui de la tâche b, sans que nécessairement le processus mental nécessaire à l'exécution de la tâche b se retrouve dans celle a.

La condition (1) peut s'exprimer par l'une des deux conditions suivantes qui sont équivalentes

$$n(a \wedge \bar{b}) = 0 \qquad (2)$$

$$n(\bar{a} \wedge b) = [n(b) - n(a)] \quad ; \quad (3)$$

ou encore

$$n(a) < n(b) \text{ et } n(a \wedge \bar{b}) = \min\{\text{card}[X \cap E(\bar{b})]/X \subset E, \text{card}(X) = n(a)\} \qquad (2')$$

$$n(a) < n(b) \text{ et } n(\bar{a} \wedge b) = \min\{\text{card}[X^c \cap E(b)]/X \subset E, \text{card}(X) = n(a)\} \qquad (3')$$

J. Loevinger (cf [18]) est partie de la condition (2), la plus directement perceptible, et a construit un indice, dont le maximum est 1 dans le cas où la condition (2) est remplie, 0 dans le cas de l'indépendance, mais qui peut prendre des valeurs négatives dans certaines situations.

Reprenons ici le tableau (10) du paragraphe précédent, mais où on remplace les cardinaux n(i,j) par les proportions p(i,j)=n(i,j)/n :

		1	0	
		$E(b)$	$E(\bar{b})$	
1	$E(a)$	$p(a \wedge b)$	$p(a \wedge \bar{b})$	$p(a)$
0	$E(\bar{a})$	$p(\bar{a} \wedge b)$	$p(\bar{a} \wedge \bar{b})$	$p(\bar{a})$
		$p(b)$	$p(\bar{b})$	

(4)

Cet indice peut être présenté sous l'une des deux formes :

$$H(a,b) = [\frac{p(a \wedge b)}{p(a)} - p(b)]/[1-p(b)] \qquad (5)$$

$$= 1 - \frac{p(a \wedge \bar{b})}{p(a)p(\bar{b})} \qquad (6)$$

La valeur de cet indice a pour objet de déceler, par rapport à l'hypothèse d'indépendance entre a et b où cette valeur est nulle, le degré d'implication:a => b, reflété par le "degré de l'inclusion $E(a) \subset E(b)$". La valeur 1 est atteinte lorsqu'on a exactement $E(a) \subset E(b)$.

Cependant, il est difficile de se rendre compte dans l'absolu, surtout lorsqu'on a à comparer plusieurs couples de comportements, à quoi correspond un degré d'implication exprimé par une valeur donnée de cet indice. Ce problème est senti de façon cruciale lorsqu'il s'agit de retenir un seuil h_o à partir duquel on décide

$$(a => b) <=> H(a,b) \geq h_o \qquad (7)$$

en attribuant aux "fluctuations d'échantillonnage" la non-réalisation stricte de $E(a) \subset E(b)$. Dans ces conditions quelle valeur h_o voisine de 1 retenir: 0,9 ? 0,8 ? 0,7 ? ...

D'où l'idée (cf.[7], R. Gras) de procéder conformément à la démarche générale de I.C. Lerman dans l'élaboration d'un indice de proximité entre attributs (cf. § II. ci-dessus) pour évaluer le caractère plus ou moins négligeable du cardinal de l'ensemble des individus contredisant l'assertion a=>b, c'est-à-dire $n(a \wedge \bar{b})$ qui jouera le rôle de l'indice brut. Mais ici, si $q(a,\bar{b})$ est l'indice centré et réduit par rapport à une h.a.l. respectant d'une certaine façon les caractéristiques de cardinalité, on ira d'autant plus dans le sens de l'implication (a => b) que $q(a,\bar{b})$ est plus fortement négatif, que $\phi[q(a,\bar{b})]$ est plus petit (ϕ étant la f.r. de la loi normale $\mathcal{N}(0,1)$), que $\psi[q(a,\bar{b})] = 1 - \phi[q(a,\bar{b})]$ est grand.

C'est donc finalement $\psi[q(a,\bar{b})]$ qui évaluera l'intensité de l'implication a => b. Le seuil permettant de retenir sur l'ensemble A des comportements une relation d'ordre, généralement partielle, sera directement défini à partir de ψ sous la forme :

$$a\ R\ b \iff \mathrm{card}(E(a)) \lesssim \mathrm{card}(E(b)) \text{ et } \psi[q(a,\bar{b})] \gtrsim \psi_o$$

pour tout (a,b) de A×A. Cette relation d'ordre sera représentée par un graphe orienté dit "graphe d'implication" (cf. [7], R. Gras).

Cette mesure de l'intensité d'implication suppose l'approximation normale de la loi de la v.a. associée à $n(a \wedge \bar{b})$. Cette approximation normale peut devenir fragile dès que $n(a \wedge \bar{b})$ devient par "trop petit" (e.g. $n(a \wedge \bar{b}) = 0,1,2$ ou 3) , ce qui, paradoxalement peut aller dans le sens de l'implication. Cet inconvénient statistique qu'il convient sans doute d'analyser plus finement, reste d'effet limité. Il se présentera beaucoup moins dans le cas du deuxième indice d'implication que nous proposerons.

L'h.a.l. qui avait jusqu'à ce travail été retenue est à caractère binomial (cf.[7]) où l'indice $q(a,\bar{b})$ considéré se met sous la forme :

$$q_b(a,\bar{b}) = \sqrt{n}[p(a \wedge \bar{b}) - p(a)p(\bar{b})] / \sqrt{p(a)p(\bar{b})[1-p(a)p(\bar{b})]} \qquad (8)$$

Mais nous avons vu au paragraphe précédent que le modèle binomial est en quelque sorte intermédiaire entre le modèle hypergéométrique et celui poissonnien qui est de toute façon une bonne approximation du modèle binomial (cf.[5]). C'est donc par rapport aux modèles 1 (hypergéométrique) et 3 (poissonnien) que nous situerons notre analyse. Le premier modèle nous donnera une forme par trop symétrique de la mesure de l'implication alors que le troisième permettra de préciser deux formes différentes d'une mesure orientée de l'implication, ayant chacune un intérêt propre et permettant de déceler plus finement une implication et de l'interpréter.

Terminons cette introduction en remarquant que l'indice H(a,b) de J. Loevinger peut également se mettre sous l'une des deux formes suivantes :

$$H(a,b) = [p(a \wedge b) - p(a)p(b)] / p(a)p(\bar{b}) \qquad (6)$$

$$H(a,b) = [p(a \wedge b)p(\bar{a} \wedge \bar{b}) - p(a \wedge \bar{b})p(\bar{a} \wedge b)] / p(a)p(\bar{b}) \qquad (7)$$

Ainsi, au facteur $\sqrt{n}$ près, le numérateur de l'indice H(a,b) correspond à l'indice centré associé à l'une des cases (1,0) ou (0,1) ; son dénominateur est le carré de celui de l'indice $q_p(a,\bar{b})$ centré réduit par rapport au modèle poissonnien.

III.2. Indices d'implication pour chacun des deux modèles hypergéométrique et poissonnien.

Partons de l'indice brut $n(a \wedge \bar{b})$ considéré dans [6] , cardinal de sous-ensemble des sujets contredisant l'implication a => b et contenu de la case (1,0) du tableau (10) (§ II précédent). L'indice centré réduit par rapport à l'h.a.1. N_1 à caractère hypergéométrique se met sous la forme

$$q_h(a,\bar{b}) = \sqrt{n-1}\,\frac{[n(a\wedge\bar{b})n(\bar{a}\wedge b) - n(a\wedge b)n(\bar{a}\wedge\bar{b})]}{\sqrt{n(a)n(\bar{a})\ n(b)n(\bar{b})}} \qquad (8)$$

(cf. formule (9) § II)

ou encore, en confondant (n-1) et n

$$q_h(a,\bar{b}) = \sqrt{n}\ \frac{[p(a\wedge\bar{b})p(\bar{a}\wedge b) - p(a\wedge b)p(\bar{a}\wedge\bar{b})]}{\sqrt{p(a)\ p(\bar{a})\ p(b)\ p(\bar{b})}} \qquad (9)$$

Comme nous l'avons déjà fait remarquer l'indice obtenu est exactement le même si au lieu de partir de l'indice brut $n(a \wedge \bar{b})$, on partait de celui $n(\bar{a} \wedge b)$, contenu de la case (0,1) du tableau (10) (§ II) ; en d'autres termes

$$q_h(\bar{a},b) = q_h(a,\bar{b}) \qquad (10)$$

Relativement au problème posé et dans le cadre de l'h.a.1. N_1 où n(a) et n(b) sont fixés, il revient tout à fait au même d'évaluer la relative petitesse de $n(a \wedge \bar{b})$ ou de $n(\bar{a} \wedge b)$ (cf. expressions (2') et (3')).

On a d'autre part vu que

$$q_h(\bar{a},b) = q_h(a,\bar{b}) = -\,q_h(a,b) = -q_h(\bar{a},\bar{b}) \qquad (10)$$

Ainsi, l'indice $q_h(a,\bar{b})$ (resp. $q_h(\bar{a},b)$) est un indice d'"opposition" entre a et b. Toutefois, il prend le sens d'un indice d'implication dans le cas où n(a)< n(b) , l'"opposition" correspondant alors à une "absorption de a par b". D'ailleurs sin(a)≥ n(b), il n'est pas question de considérer a => b. D'autre part, on peut remarquer que par rapport à l'indice $q_h(a,b)$, celui (7) H(a,b) de J. Loevinger accentue le rôle de la "petitesse" de n(a) et de la "grosseur" de n(b).

Si on se réfère au modèle poissonnien de l'h.a.1., l'indice centré réduit (avant référence à une échelle de probabilité) qui se rapproche le plus dans son expression de celui de J. Loevinger, est celui $q_p(a,\bar{b})$ associé à $n(a \wedge \bar{b})$, contenu de la case (1,0) de la table de contingence 2x2 ; il s'agit

exactement de

$$q_p(a,\bar{b}) = \frac{n(a \wedge \bar{b}) - [n(a)n(\bar{b})/n]}{\sqrt{n(a)n(\bar{b})/n}}$$

$$= \sqrt{n} \times \frac{[p(a \wedge \bar{b}) - p(a)p(\bar{b})]}{\sqrt{p(a)p(\bar{b})}} \qquad (11)$$

$$= \sqrt{n} \times \frac{[p(a \wedge \bar{b})p(\bar{a} \wedge b) - p(a \wedge b)p(\bar{a} \wedge \bar{b})}{\sqrt{p(a)p(\bar{b})}}$$

Mais l'indice H(a,b), dans la comparaison de plusieurs couples de comportements, force encore davantage que ne le fait $q_p(a,\bar{b})$ sur la concomitance de la rareté de a et de la fréquence de b. Il s'agit d'une qualité relative, car de la sorte on accentue la valeur de l'implication pour des situations où n(a) est "petit" et n(b) "grand" qui sont certes compatibles avec la raison (i), mais aussi tout à fait conformes avec la raison (ii) (cf. début du paragraphe III.1. ci-dessus).

Cependant, on a pu voir que pour n(a) et n(b) donnés tels que $n(a) < n(b)$, pour se rendre compte du degré d'implication de a => b reflété par le "degré d'inclusion de E(a) dans E(b)", on peut (cf. expression (3') par rapport à (2') ci-dessus) plutôt partir de $n(\bar{a} \wedge b)$ que de $n(a \wedge \bar{b})$; ces deux approches coïncident d'ailleurs dans le cadre du modèle hypergéométrique. En partant de $n(\bar{a} \wedge b)$, on évite en général l'inconvénient statistique de partir d'un indice brut trop faible rendant sujette à caution l'approximation normale permettant de passer de l'indice centré réduit à une échelle uniforme de probabilité pour la mesure de l'intensité de l'implication. Toutefois, cette raison n'a rien de fondamental dans la comparaison des deux indices ; le deuxième qui nous était apparu dans l'analyse du premier, a pour expression

$$q_p(\bar{a},b) = \frac{n(\bar{a} \wedge b) - [n(\bar{a})n(b)/n]}{\sqrt{n(\bar{a})n(b)/n}}$$

$$= \frac{\sqrt{n}\,[p(\bar{a} \wedge b) - p(\bar{a})p(b)]}{\sqrt{p(\bar{a})p(b)}} \qquad (12)$$

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

Une autre différence importante entre l'indice de J. Loevinger H(a,b) et les deux indices $q_p(a,\bar{b})$ et $q_p(\bar{a},b)$ que nous venons de présenter, concerne l'influence de la taille n de l'échantillon définissant l'ensemble E. L'indice H(a,b) est invariant par une dilatation de l'ensemble E et des sous-ensembles E(a) et E(b), qui préserve les proportions p(a), p(b) et p(a∧b). Par contre, chacun de nos deux indices augmente proportionnellement à $\sqrt{n}$. Ainsi ces indices tiennent heureusement compte de la taille de l'échantillon ; pour une même position relative de E(a) et de E(b) dans E, la relation d'implication (a => b) est d'autant plus marquée que la représentativité de l'échantillon augmente.

III.3. Comparaison des deux indices $q_p(a,\bar{b})$ et $q_p(\bar{a},b)$ Graphes d'implication associés

α) Transitivité ; inégalité triangulaire

Si les intensités de chacune des deux implications a => b et b => c sont fortes , on peut s'attendre à ce que l'intensité de l'implication a => c soit appréciable. C'est exactement ce que traduiront les inégalités (16) et (17) ci-dessous, directement obtenues à partir de l'inégalité triangulaire, respectivement pour chacun des deux indices.

Le cardinal de la différence symétrique définit une distance sur l'ensemble des parties d'un ensemble ; il en résulte dans notre contexte l'expression suivante de l'inégalité triangulaire pour cette distance :

$$[n(a\wedge\bar{c})+n(\bar{a}\wedge c)]\lesssim[n(a\wedge\bar{b})+n(\bar{a}\wedge b)]+[n(b\wedge\bar{c})+n(\bar{b}\wedge c)] \qquad (13)$$

Compte tenu des relations de la forme

$$n(\bar{a}\wedge c)+n(a)=n(a\wedge\bar{c})+n(c) \qquad (14),$$

la relation (13) donne les deux relations suivantes

$$\begin{aligned} n(a\wedge\bar{c}) &\lesssim n(a\wedge\bar{b})+n(b\wedge\bar{c}) \\ n(\bar{a}\wedge c) &\lesssim n(\bar{a}\wedge b)+n(\bar{b}\wedge c) \end{aligned} \qquad (15)$$

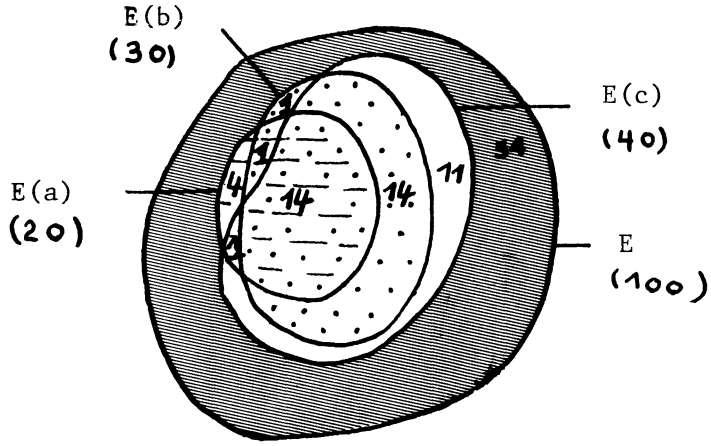
qui donnent directement les deux relations suivantes

$$q_p(a,\bar{c}) \lesssim \sqrt{\frac{n(\bar{b})}{n(\bar{c})}}\ q_p(a,\bar{b}) + \sqrt{\frac{n(b)}{n(a)}}\ q_p(b,\bar{c}) + \frac{1}{\sqrt{\mu(a,\bar{c})}}\ [\mu(a,\bar{b}) + \mu(b,\bar{c}) - \mu(a,\bar{c})], \quad (16)$$

$$q_p(\bar{a},c) \lesssim \sqrt{\frac{n(b)}{n(c)}}\ q_p(\bar{a},b) + \sqrt{\frac{n(\bar{b})}{n(\bar{a})}}\ q_p(\bar{b},c) + \frac{1}{\sqrt{\mu(\bar{a},c)}}\ [\mu(\bar{a},b) + \mu(\bar{b},c) - \mu(\bar{a},c)] \quad (17)$$

où μ (e,f) désigne n(e)n(f)/n , e et f indiquant deux attributs quelconques.

Terminons en illustrant chacune des deux inégalités sur l'exemple suivant où



n=100, n(a)=20, n(b)=30, n(c)=40 ;
$n(a \wedge \bar{b})=5$, $n(a \wedge \bar{c})=5$, $n(b \wedge \bar{c})=2$;
$n(\bar{a} \wedge b)=15$, $n(\bar{a} \wedge c)=25$ et $n(\bar{b} \wedge c)=12$.

On obtient
$q_p(a,\bar{b})= -2,41$ et $q_p(b,\bar{c})= -3,77$. L'inégalité (16) précédente donne $q_p(a,\bar{c}) \lesssim -1,45$, alors que dans l'exemple $q_p(a,\bar{c})=-2,02$. D'autre part $q_p(\bar{a},b)=-1,84$ et $q_p(\bar{b},c) =-3,02$. L'inégalité (17) précédente donne $q_p(\bar{a},c) \lesssim -0,88$, alors que dans l'exemple $q_p(\bar{a},c)=-1,24$.

On se rend compte que les inégalités (16) et (17) peuvent avoir un intérêt dans la reconnaissance du degré d'implication a => c dès que les intensités de chacune des deux implications a => b et b => c sont assez fortes.

β) Sens de variation

Nous allons ici chercher à détecter le sens de variation de chacun des deux indices d'implication $q_p(a_o,\bar{b})$ et $q_p(\bar{a}_o,b)$ par rapport à certains paramètres fixant la position de E(b) relativement à $E(a_o)$, lorsque $E(a_o)$ étant fixé, E(b) varie d'une "certaine façon" en augmentant de cardinal.

Les numérateurs de chacun des deux indices $q_p(a_o,\bar{b})$ et $q_p(\bar{a}_o,b)$ sont égaux. D'où l'idée de neutraliser l'effet du numérateur en étudiant les sens de variation de chacun des deux indices sur l'ensemble des parties E(b) pour lesquels le numérateur commun des indices reste fixé ; soit

$$\{E(b)/E(b) \subset E \text{ et } \left[n(a_o \wedge b) - \frac{n(a_o)n(b)}{n}\right] = \nu\} \qquad (18)$$

où $(-\nu)$ est la valeur commune des numérateurs des deux indices précédents.

L'ensemble (18) précédent n'est pas vide s'il existe un couple d'entiers positifs h et k tels que

$$h \leq n(a_o) \wedge k \text{ et } h - \frac{n(a_o)k}{n} = \nu \qquad (19)$$

où h joue le rôle de $n(a_o \wedge b)$ et k celui de n(b). Il y a dans ces conditions $\binom{n(a_o)}{h}\binom{n-n(a_o)}{k-h}$ parties E(b) qui réalisent

$$n(a_o \wedge b) = h \text{ et } n(b) = k$$

Pour fixer les idées, soit un tel couple d'entiers (h_o,k_o) à partir duquel nous allons montrer comment construire l'ensemble (18) ci-dessus.

Propriété 1. $M(a_o)$ désignant un multiple entier positif commun à $n(a_o)$ et à n , il y a

$$\binom{n(a_o)}{m+h_o}\binom{n-n(a_o)}{l-m+k_o-h_o} \qquad (20)$$

éléments de l'ensemble (18) de parties pour lesquels

$$n(b) = 1+k_o \text{ et } n(a_o \wedge b) = m+h_o$$

où $1 = M(a_o)/n(a_o)$ et $m = M(a_o)/n$.

En effet, en supposant

$$\begin{aligned} &m + h_o \leqslant n(a_o) \\ &\text{et} \\ &1 + k_o \leqslant n - \{n(a_o)-(m+h_o)\} \end{aligned} \qquad (21)$$

si on pose

$$\begin{aligned} &n(b) = 1 + k_o \\ &\text{et} \quad n(a_o \wedge b) = m + h_o \quad , \end{aligned} \qquad (22)$$

on vérifie aisément que

$$n(a_o \wedge b) - [n(a_o)n(b)/n] = h_o - \frac{n(a_o)k_o}{n} = \nu \qquad (23)$$

Dans ces conditions, à partir de la plus petite valeur k_o de k, à laquelle se trouve associée h_o, on imaginera de faire "grossir" l'ensemble E(b) à partir de la suite croissante des multiples communs de $n(a_o)$ et de n, en s'assurant à chaque fois des conditions (21).

Illustrons la construction précédente en partant d'un couple $(E(a_o),E(b_o))$ de parties d'un ensemble E de cardinal n=100, où $\text{card}(E(a_o)) = n(a_o) = 10$, $\text{card}(E(b_o)) = 20$ et $\text{card}(E(a_o) \cap E(b_o)) = n(a_o \wedge b_o) = 8$. En codant E par l'ensemble des entiers {00,01,02,.....,98,99}, on peut prendre par exemple

$E(a_o)$ = {00,01,02,03,04,05,06,07,08,09}

et $E(b_o)$ = {00,01,02,03,04,05,06,07,10,11,12,...,20,21}

Le premier multiple commun de $n(a_o)=10$ et de n=100 est $M(a_o)=100$ auquel il lui correspond, avec les notations de l'énoncé précédent, 1=10 et m=1. D'où $n(b) = 1+n(b_o)=30$ et $n(a_o \wedge b) = m+n(a_o \wedge b_o)=9$.

On peut en particulier produire, parmi les $\binom{10}{9}\binom{90}{21}$ ensembles E(b) possibles, celui $E(b_1)$ suivant :

$E(b_1)= \{00,01,02,03,04,05,06,07,08,10,11,\ldots,28,29,30\}$

où on a ici $E(b_o) \subset E(b_1)$.

Propriété 2. Pour $E(a_o)$ et $\nu>0$ fixés, les comportements des deux fonctions $\psi[q_p(a_o,\bar{b})]$ et $\psi[q_p(\bar{a}_o,b)]$ (cf. § III.1.) sont opposés sur l'ensemble (18) des parties E(b) :

$\psi[q_p(a_o,\bar{b})]$ est une fonction croissante de

$$n(b),\ n(\bar{a}_o \wedge b)/n(b) \text{ et } n(\bar{a}_o \wedge \bar{b})/n(\bar{b}) \quad ;$$

décroissante de (24)

$$n(a_o \wedge b)/n(b) \text{ et } n(a_o \wedge \bar{b})/n(\bar{b})$$

Au contraire $\psi[q_p(\bar{a}_o,b)]$ est une fonction décroissante de

$$n(b)\ ,\ n(\bar{a}_o \wedge b)/n(b) \text{ et } n(\bar{a}_o \wedge \bar{b})/n(\bar{b}) \quad ;$$

croissante de (25)

$$n(a_o \wedge b)/n(b) \text{ et } n(a_o \wedge \bar{b})/n(\bar{b}).$$

Ces résultats sont immédiats à partir des relations suivantes :

$$\begin{cases} \dfrac{n(a_o \wedge b)}{n(b)} = \dfrac{\nu}{n(b)} + \dfrac{n(a_o)}{n} & \text{(fonction décroissante de } n(b)) \\ \dfrac{n(a_o \wedge \bar{b})}{n(\bar{b})} = \dfrac{-\nu}{n(\bar{b})} + \dfrac{n(a_o)}{n} & \text{(fonction décroissante de } n(b)) \\ \dfrac{n(\bar{a}_o \wedge b)}{n(b)} = \dfrac{-\nu}{n(b)} + \dfrac{n(\bar{a}_o)}{n} & \text{(fonction croissante de } n(b)) \\ \dfrac{n(\bar{a}_o \wedge \bar{b})}{n(\bar{b})} = \dfrac{\nu}{n(\bar{b})} + \dfrac{n(\bar{a}_o)}{n} & \text{(fonction croissante de } n(b)). \end{cases} \quad (26)$$

Il ne faut pas perdre de vue que les deux indices restent très "parallèles" dans leur comportement global à travers tous les couples d'items (a_o,b) pour lesquels $n(a_o) \leq n(b)$ et $\nu(a_o,b) > 0$. Les comportements inverses des deux indices $\psi[q_p(a_o,\bar{b})]$ et $\psi[q_p(\bar{a}_o,b)]$ ont un caractère tout à fait local puisque la variation est étudiée à $\nu(a_o,b)$ constant. Chacun des deux indices met l'accent sur un aspect de la perception du degré de l'inclusion de $E(a_o)$ dans E(b).

Il est en effet intuitivement souhaitable que le degré d'implication soit une fonction décroissante du premier (resp. croissante du second) de chacun des deux paramètres $n(a_o \wedge \bar{b})/n(\bar{b})$ et $n(\bar{a}_o \wedge \bar{b})/n(\bar{b})$ qui sont directement liés au cardinal de l'ensemble $E(a_o) \cap E(\bar{b})$ formé des individus violant, au sens logique du terme, l'implication $a_o \Rightarrow b$.

Toutefois, il est également intuitivement désirable que le degré d'implication soit une fonction décroissante (resp. croissante) de n(b) et $n(\bar{a}_o \wedge b)/n(b)$ (resp. de $n(a_o \wedge b)/n(b)$), car de la sorte, on insiste sur le caractère exceptionnel de l' "absorption" de a_o par b. Pour $\nu(a_o,b)$ positif fixé, le premier indice $\psi[q_p(a_o,\bar{b})]$ va dans le sens de la première condition , mais à l'encontre de la seconde. Inversement, le second indice $\psi[q_p(\bar{a}_o,b)]$ va dans le sens de la deuxième condition, mais à l'encontre de la première.

Revenons un instant sur les conditions (i) et (ii) introduites au début du paragraphe III.1. Relativement à un couple de comportements (a,b) dans un contexte donné, il est en général très difficile de séparer chacun des deux facteurs respectivement sous-jacents à (i) et à (ii). En effet, ces derniers interfèrent puisque le premier facteur contribue directement au degré de la complexité. D'autre part, ces deux facteurs contribuent également à l'élévation de $\nu(a,b) = n(a \wedge b) - [n(a)n(b)/n]$. Toutefois, en raison des propriétés locales précédentes, nous croyons l'indice $\psi[q_p(a,\bar{b})]$ d'abord sensible au non respect de l'implication, au sens logique du terme, alors que $\psi[q_p(\bar{a},b)]$ peut mieux mettre en relief les faibles variations de la complexité, comme d'ailleurs l'analyse didactique (cf. § IV) le montrera.

γ) Graphes d'implication

Chacun des deux indices d'implication $\psi[q_p(a,\bar{b})]$ et $\psi[q_p(\bar{a},b)]$ permet la définition d'un graphe orienté, valué et sans circuits sur l'ensemble A des stimuli, défini comme suit : Pour tout couple (a,b) d'éléments distincts de A, il existe un arc orienté d'origine a et d'extrémité b, si et seulement si card[E(a)] < [card E(b)] ; d'autre part, la valuation portée sur un tel arc (a,b) est définie par la valeur de l'indice d'implication choisi.

Si on ne fait pas rentrer en ligne de compte la valuation, le graphe est celui d'un préordre total sur A ; une même classe du préordre est définie par l'ensemble des items dont la fréquence de réussite est la même; elle peut donc être exprimée sous la forme

$$\{a \in A/\ \mathrm{card}[E(a)] = k\} \qquad (27)$$

Comme nous l'avons exprimé au paragraphe III.1, du graphe pondéré on passe à un graphe discret par l'adoption d'un seuil ψ_o. On peut ainsi définir, relativement à chacun des deux indices d'implication, un graphe orienté et sans circuits qu'on appellera "graphe d'implication". De façon explicite, le premier (A,U_1) où U_1 désigne l'ensemble des arcs, sera défini par :

$$(\forall (a,b) \in AxA),\ (a,b) \in U_1 \iff \begin{cases} E(a) \subset E(b), \text{ ou bien} \\ \mathrm{card}(E(a)) < \mathrm{card}(E(b)) \text{ et} \\ \psi[q_p(a,\bar{b})] \geq \psi_o \end{cases} \qquad (28)$$

Le second graphe (A,U_2) où U_2 désigne l'ensemble des arcs, sera défini par :

$$(\forall (a,b) \in A \times A),(a,b) \in U_2 \iff \begin{cases} E(a) \subset E(b), \text{ ou bien} \\ \mathrm{card}(E(a) < \mathrm{card}(E(b)) \text{ et} \\ \psi[q_p(\bar{a},b)] \geq \psi_o \end{cases} \qquad (29)$$

Si on se réfère à la pratique des tests statistiques où on adopte un seuil de signification $\alpha = 0{,}05$, il y a lieu de prendre $\psi_o = 0{,}95$. En substituant de cette manière le graphe discrétisé au graphe pondéré, nous passons de notre approche en analyse des données, où l'h.a.l. joue le rôle d'une hypothèse de référence qui fournit l'échelle de mesure de l'intensité d'une relation en termes de vraisemblance, à une optique décisionnelle plus classique.

Propriété 3. Le graphe (A,U_2) est contenu dans celui (A,U_1).

On le voit immédiatement ; en effet, $q_p(\bar{a},b) < q_p(a,\bar{b})$ dès que $n(a) < n(b)$, pour tout couple (a,b) de $A \times A$.

On peut ainsi considérer que le graphe (A,U_2) est celui des implications les plus "fortes". Toutefois, ce qu'il gagne en force il le perd en richesse, notamment par rapport à (A,U_1) qui peut comprendre beaucoup plus d'arcs.

Pour avoir la même base de comparaison et pouvoir ainsi distinguer les tendances naturelles de chacun des deux indices , nous nous sommes inspirés de notre pratique de la réduction globale des similarités en classification

hiérarchique par l'algorithme de la vraisemblance des liens (cf. fin du paragraphe II.5). Dans ces conditions, on commence par distinguer l'ensemble des couples suivants

$$G = \{(a,b) \in A \times A / \mathrm{card}(E(a)) < \mathrm{card}(E(b)), E(a) \cap E(\bar{b}) \neq \emptyset\} \qquad (30)$$

sur lequel on réduit globalement chacun des deux indices $q_p(a,\bar{b})$ et $q_p(\bar{a},b)$. En d'autres termes on remplace $q_p(a,\bar{b})$ et $q_p(\bar{a},b)$ respectivement par

$$q'_p(a,\bar{b}) = [q_p(a,\bar{b}) - \bar{q}_{10}]/\tau_{10}$$

et

$$q'_p(\bar{a},b) = [q_p(\bar{a},b) - \bar{q}_{01}]/\tau_{01} \qquad (31)$$

pour tout couple (a,b) de G, où $\bar{q}_{10}$ et τ^2_{10} (resp. $\bar{q}_{01}$ et τ^2_{01}) sont la moyenne et la variance de la distribution sur G de $q_p(a,\bar{b})$ (resp. $q_p(\bar{a},b)$).

Chacun des deux graphes (28) et (29) sera alors respectivement défini en remplaçant $q_p(a,\bar{b})$ par $q'_p(a,\bar{b})$ et $q_p(\bar{a},b)$ par $q'_p(\bar{a},b)$, pour tout (a,b) de G.

Comme nous l'annoncions dans l'introduction, nous allons dans un prochain article, représenter dans un cas réel chacun des deux graphes d'implication (28) et (29) où, pour avoir une même base de comparaison, nous substituerons $q'_p(\bar{a},b)$ à $q_p(a,\bar{b})$. L'interprétation comparée des deux graphes permettra de se rendre compte de l'intérêt de chacun des deux indices.

Nous chercherons ensuite à évaluer globalement chacun des deux graphes pondéré ou discret, associé à chacun des deux indices, par rapport à l'hypothèse de la taxinomie d'objectifs cognitifs qui est reflétée par un préordre total sur l'ensemble A des comportements. Cette évaluation se fera au moyen d'un indice très général de comparaison entre deux relations pondérées sur un ensemble fini. Ce dernier indice peut jouer le rôle de critère d'adéquation.

(à suivre)

BIBLIOGRAPHIE

[1] BLOOM B. et collaborateurs, *Taxinomie des objectifs pédagogiques*, Tome 1 : *domaine cognitif*, Montréal, Edition Nouvelle, 1969.

[2] BROUSSEAU G., "Evaluation et théorie de l'apprentissage en situation scolaire", *Conférence de Campinas (Brésil), février 1979*, texte ronéoté, I.R.E.M. de Bordeaux.

[3] DEGENNE A., *Techniques ordinales en analyse des données statistiques*, Tome II, Paris, Hachette, 1972.

[4] FELLER W., *An introduction to probability theory and its applications*, Volume I, second edition, New York, John Wiley, 1964.

[5] FLAMENT C., DEGENNE A., VERGES P., "Analyse de similitude ordinale" *in Actes du Colloque International DGRST-CNRS, Marseille 11-13 Décembre 1975, Informatique et Sciences Humaines, n 40-41, Mars/Juin 1979.*

[6] GRAS R., *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions congnitives et de certains objectifs didactiques en mathématiques*, Thèse d'état, Université de Rennes-I, 1979.

[7] GUTTMAN L., STOUFFER S.A., SUCHMAN E.A., LAZARSFELD P.F., *The American Soldier, Vol.4 : Measurement and Prediction*, Princeton, University Press, 1950.

[8] LECALVE G., *Problèmes d'analyse des données*, thèse d'état (2ème partie), Université de Rennes-I, 1976.

[9] LERMAN I.C., "Combinatorial analysis in the statistical treatment of behavioral data", *Quality and Quantity* , 14 , 1980, 431-469.

[10] LERMAN I.C."Formal analysis of a general notion of proximity between variables", *Actes partiels du colloque "Congrès Européen des Statisticiens", Grenoble, septembre 1976.*

[11] LERMAN I.C., "Etude formelle et statistique de la notion de ressemblance", *rapport de recherche I.R.I.S.A.*, n°107, Décembre 1979.

[12] LERMAN I.C., "Croisement de classifications floues", *Publ. Inst. Stat. Univ. Paris*, 1979, XXIV, fasc. 1-2, 13-46 .

[13] LERMAN I.C., "Analyse ordinale d'une classe d'échelles" *in Analyse des données, tome I*, Paris, Publications A.P.M.E.P., 1981, 40 p.

[14] LOEVINGER J., "A systematic approach to the construction and evaluation of tests of ability", *Psychological Monographs*, 61, n°4, 1947 .

[15] MATALON B., *L'analyse hiérarchique*, Paris, La Haye, Mouton-Gauthier-Villars, 1965.

[16] TOURNEUR Y., "Classification des questions d'évaluation en mathématique" et "Taxonomie des objectifs cognitifs en mathématique : étude du modèle N.L.S.M.A.", *Mathematica et Paedagogia*, n° 56 et 57, 1972.

[17] VERGNAUD G., "Activité et connaissance opératoire", *Bulletin A.P.M.E.P.*, n°307, février 1977.