MATHÉMATIQUES ET SCIENCES HUMAINES

RÉGIS GRAS
ANNIE LARHER

L'implication statistique, une nouvelle méthode d'analyse de données

Mathématiques et sciences humaines, tome 120 (1992), p. 5-31
<http://www.numdam.org/item?id=MSH_1992__120__5_0>

Math. Inf. Sci. hum. (31e année, n° 120, *1993*, p. 5-31).

L'IMPLICATION STATISTIQUE, UNE NOUVELLE MÉTHODE D'ANALYSE DE DONNÉES

Régis GRAS et Annie LARHER[1]

RÉSUMÉ — *Si le problème de la concomitance de deux variables a et b trouve une partie de sa réponse dans l'étude symétrique de la corrélation ou dans celle de la similarité, celui de l'implication (si a alors b) passe, en revanche, par l'examen d'une relation dissymétrique. Par rapport à une problématique psychologique de complexité, R. GRAS, dans sa thèse, a apporté une contribution qui a permis de nombreuses applications de ce type de relation dans des travaux de recherche en psychologie génétique et en didactique des mathématiques, domaines non exclusifs d'autres champs d'application. Mais les variables considérées dans sa recherche se limitent aux variables binaires, présence-absence d'un caractère chez un individu donné. Il s'agit ici d'étendre l'étude de l'implication statistique (ou quasi-implication) à d'autres types de variables et, surtout, à des classes de telles variables. Cette extension nous permet de construire un arbre de classes orientées. Ces développements et cette construction originale résultent, principalement, de la thèse d'Annie LARHER.*

SUMMARY — Statistical inference, a new method of data analysis.

If the problem of the simultaneity of two variables a and b finds part of its solution in the symmetrical study of correlation or similarity, the question of inference (if a then b) has on the other hand to undergo the examining of a dissymetrical relationship. In relation with a psychological problem of complexity, R. GRAS brought in his thesis a contribution that made possible a number of applications of this type of relationship in research in psychological genetics or in mathematical didactics, those areas not precluding other application fields. But the variables taken into account in his research are restricted to binary ones, presence-absence of a feature in a given individual. The purpose here is to extend the study of the statistical inference (or quasi-inference) to other types of variables and above all, to classes of such variables. This extension allows us to build a tree of oriented classes. These developments and this original construction mainly proceed from Annie LARHER's thesis.

INTRODUCTION

La notion de similarité entre attributs *a* et *b* - ou variables binaires - est essentiellement symétrique. Mais s'agissant de répondre à l'étude de "*si a alors b*", R. GRAS [GRAS 1979] puis I.C. LERMAN, R. GRAS, H. ROSTAM [LERMAN, GRAS, ROSTAM, 1981] ont défini et précisé la notion de quasi-implication mesurée par une intensité d'implication et la notion de graphe d'implication, image de la relation de préordre partiel qui en découle. L'intensité d'implication s'exprime, dans le cas général, par une intégrale gaussienne, qui rend compte du caractère invraisemblable de l'observation faite du nombre de cas où l'implication $a \Rightarrow b$ se trouve défaillante, dans une hypothèse d'absence de lien implicatif a priori. Des propriétés de cette intensité, rappelées ou étudiées dans la thèse d'A. LARHER [LARHER, 1991], sous la direction de R. GRAS, seront présentées ici. Mais l'originalité de cette thèse et des travaux menés depuis consiste principalement, toujours de façon dissymétrique :

[1] Institut de Recherche Mathématique de Rennes, Campus de Beaulieu - 35042 RENNES CEDEX.

- d'une part, dans l'extension à d'autres types de variables (modales et fréquentielles) de la notion d'intensité d'implication et de son uniformisation,
- d'autre part , dans la construction nouvelle d'un indice d'implication entre classes a et b de variables, étendant l'évaluation de l'implication $a \Rightarrow b$ à celle de $\mathcal{A} \Rightarrow \mathcal{B}$, sur la base d'une cohésion implicative suffisante des classes examinées et d'une relation implicative maximale entre leurs éléments respectifs,
- enfin dans l'organisation en structure arborescente de l'ensemble de classes, organisation empruntée à la classification hiérarchique construite avec l'indice de I.C. LERMAN. Mais ici, le lien entre deux noeuds de l'arbre est orienté.

Nous examinerons plus en détail ces deux volets.

Les concepts théoriques qui s'y rattachent ne doivent pas être considérés comme des spéculations détachées de la volonté de construire des modèles de comportements ou de processus de pensée. Au contraire, ils s'inscrivent en réponse, provisoire sans doute, à quelques questions levées par notre problématique actuelle en didactique des mathématiques mais extensible à d'autres champs, questions qui serviront de support intuitif à la modélisation puis la formalisation qui suivront ; par exemple :
- à un niveau de cursus donné, peut-on, dans une situation-problème donnée, déterminer une hiérarchie partiellement ordonnée de procédures de résolution de problèmes de mathématiques, signes d'une connaissance en voie de constitution ?

- à un niveau de cursus donné, peut-on définir à partir de classes ordonnées de procédures, des conceptions homogènes et résistantes relativement à un certain savoir [2] ?
etc.

Comme nous en verrons un exemple dans le § 3.1 nous avons utilisé les modèles statistiques élaborés pour des variables (§ 1), puis des classes des variables (§ 2) afin d'outiller les utilisateurs, le didacticien en particulier, dans l'approche de questions telles que ci-dessus.

Notre problématique croise celle de certains chercheurs en intelligence artificielle dont nous reparlerons plus loin et dont le souci est :
- d'une part, la considération modale de la relation d'attribution d'un descripteur déterminé à un objet (ou un sujet) et l'apprentissage d'une base de connaissances,
- d'autre part, la reconnaissance de formes.

Cependant, comme nous venons de le voir, bien que la première problématique ne soit pas orientée par les problèmes d'apprentissage (au sens de l'I.A.) ou de structure de base de connaissances, nos points de vue pourraient se rejoindre dans la nécessité de placer les sujets, ayant conduit à une classification hiérarchique ou implicative, par rapport aux ensembles de variables classifiées.

1. IMPLICATION ENTRE VARIABLES BINAIRES ET EXTENSION

1.1. Modélisation

Dans le cas binaire, la situation générique est la suivante. Croisant une population E et un ensemble de variables V et du fait de l'observation exceptionnelle de l'implication stricte de la variable a sur la variable b, on veut donner un sens statistique à une implication non stricte :

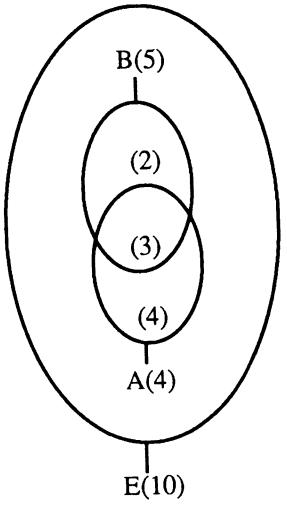
---

[2] Citons, par exemple, les conceptions "causaliste" ou "chronologiste" que de jeunes étudiants ont de la notion de probabilité conditionnelle.

$a \Rightarrow b$. En termes ensemblistes, A et B représentant les sous-populations respectives où les variables a et b prennent la valeur 1 (ou "*vrai*"), il y a équivalence à mesurer l'inclusion non stricte de A dans B.

Par exemple, si a est l'attribut "*cheveux blonds*" et b l'attribut "*yeux bleus*" dans une population E d'étudiants, les données pour étudier si $a \Rightarrow b$, dans le cas où Card E = 10, peuvent se présenter de 3 façons différentes :

V Sujets	a	b
1	0	0
2	0	1
3	1	1
4	1	0
5	0	0
6	1	1
7	1	1
8	0	0
9	0	1
10	0	0
Total	4	5



b a	1	0	Marges
1	3	1	4
0	2	4	6
Marges	5	5	10

S'inspirant de la méthode de I.C. LERMAN [81] pour définir la similarité, R. GRAS [79] axiomatise la notion d'implication statistique de la façon suivante :

Soit X et Y deux parties aléatoires quelconques de E, choisies indépendamment (absence de lien a priori) et de mêmes cardinaux respectifs que A et B. Soit $\overline{Y}$ et $\overline{B}$ les complémentaires respectifs de Y et de B dans E.

AXIOME 1

(a $\Rightarrow$ b) *est admissible au niveau de confiance* $1 - \alpha$ *si et seulement si*

$$Pr\left[card\left(X \cap \overline{Y}\right) \leq card\left(A \cap \overline{B}\right)\right] \leq \alpha$$

Intuitivement et qualitativement, ceci signifie que l'implication a $\Rightarrow$ b sera admissible à l'issue d'une expérience si le nombre d'individus de E la contredisant dans l'expérience est invraisemblablement petit par rapport au nombre d'individus attendu dans une hypothèse d'absence de lien. Par exemple si Card E = 100, Card A = 35, Card B = 50, alors Card $(A \cap \overline{B})$ = 3 est "*invraisemblablement petit*" pour une absence de lien entre a et b.

La modélisation probabiliste que nous retenons de façon privilégiée est décrite par un processus de tirage aléatoire en 3 étapes (cf. I.C. LERMAN, R. GRAS, H. ROSTAM [81]).

Notons $n_a, n_b, n_{\overline{b}}, n_{a \wedge \overline{b}}, n_{a \vee \overline{b}}$ les cardinaux respectifs de A , B , $\overline{B}$, $A \cap \overline{B}$, $A \cup \overline{B}$:

- on considère le référentiel E comme la réalisation d'un référentiel aléatoire $\mathcal{E}$ dont le cardinal $\mathcal{N}$ serait une variable aléatoire de Poisson de paramètre le cardinal n de E observé, hypothèse compatible avec les situations les plus fréquemment rencontrées et qui ne nuit pas à la généralité de la modélisation :

$$Pr\left[\mathcal{N} = m\right] = \frac{n^m}{m!} e^{-n}$$

- le choix aléatoire d'une partie quelconque (par exemple X) de cardinal aléatoire K pour une distribution uniforme de probabilité sur les éléments de $\mathfrak{C}$ et égale à $\frac{d}{m}$ (dans le cas de X, $d = n_a$) est alors de type binomial :

$$\Pr[K = k / \mathfrak{N} = m] = \binom{m}{k} \left(\frac{d}{m}\right)^k \left(1 - \frac{d}{m}\right)^{m-k} \text{(pour } k \leq m\text{)}$$

- X et $\overline{Y}$ étant deux parties quelconques choisies de façon indépendante parmi les parties ayant respectivement pour cardinaux n_a et $n_{\bar{b}}$, la probabilité qu'un élément de $\mathfrak{C}$ appartienne à $X \cap \overline{Y}$ est :

$$p(a)\,p(\bar{b}) \text{ où } p(a) = \frac{n_a}{m} \text{ et } p(\bar{b}) = \frac{n_{\bar{b}}}{m}.$$

PROPOSITION 1.
La variable aléatoire Card $(X \cap \overline{Y})$ *suit la loi de Poisson de paramètre* $n\,p(a)\,p(\bar{b})$.

La loi de probabilité conditionnelle du cardinal de $X \cap \overline{Y}$ est binomiale de paramètres m *et* $\pi = p(a)\,p(\bar{b})$

$$\Pr\left[\text{Card}(X \cap \overline{Y}) = s \,/\, n = m\right] = \binom{m}{s} \pi^s (1 - \pi)^{m-s}$$

pour $s \leq n_{a \wedge \bar{b}}$ *et* $m \geq n_{a \vee \bar{b}}$

Par suite, en faisant varier le conditionnement de $\mathfrak{N}$, on obtient :

$$\Pr[\text{Card}(X \cap \overline{Y}) = s] = \sum_{m \geq s} \Pr[\text{Card}(X \cap \overline{Y}) = s / \mathfrak{N} = m] \times \Pr[\mathfrak{N} = m]$$
$$= \frac{(n\pi)^s}{s!} e^{-n\pi}.$$

La variable aléatoire Card $(X \cap \overline{Y})$ suit donc la loi de Poisson de paramètre $n\,\pi = n\,p\,(a)p(\bar{b})$ (de moyenne et de variance $n\,\pi$). D'autres modélisations conduiraient à une loi hypergéométrique ou à une loi binomiale.

COROLLAIRE. Comme I.C. LERMAN, nous réduisons et centrons cette variable de Poisson en la variable :

$$Q(a,b) = \frac{Card\,(X \cap \overline{Y}) - n\,p(a)\,p(\overline{p})}{\sqrt{n\,p(a)\,p(\bar{b})}} = \frac{Card(X \cap \overline{Y}) - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a\,n_{\bar{b}}}{n}}}$$

Dans l'expérience, la valeur observée de $Q(a,\bar{b})$ est $q\,(a,\bar{b}) = \dfrac{n_{a \wedge \bar{b}} - \frac{n_a\,n_{\bar{b}}}{n}}{\sqrt{\frac{n_a\,n_{\bar{b}}}{n}}}$, indicateur de la non-implication de a sur b.

Dans les cas légitimant convenablement l'approximation $\left(\frac{n_a\,n_{\bar{b}}}{n} \geq 3\right)$, la variable $Q(a,\bar{b})$ suit la loi normale centrée réduite. *L'intensité d'implication*, qualité de l'admissibilité de $a \Rightarrow b$, pour $n_a \leq n_{\bar{b}}$, est alors définie à partir de l'indice $q(a,\bar{b})$ par :

DÉFINITION 1.

$$\varphi(a,\bar{b}) = 1 - \Pr[Q(a,\bar{b}) \leq q(a,\bar{b})]$$
$$= \frac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} e^{-t^2/2}\, dt$$

AXIOME 1'. L'axiome 1 devient :

L'implication $a \Rightarrow b$ sera admissible au niveau de confiance $1 - \alpha$ si et seulement si
$$\varphi(a,\bar{b}) = 1 - \Pr[Q(a,\bar{b}) \leq q(a,\bar{b})] \geq 1 - \alpha$$

Notons que, si l'approximation gaussienne n'est pas valide, il est loisible de revenir aux origines poissonniennes de la variable Card $(X \cap \bar{Y})$ et de considérer :

$$\varphi(a,\bar{b}) = 1 - Pr[Card(X \cap \bar{Y}) \leq n_{a\wedge\bar{b}}] .$$

Exemple,

b / a	1	0	Marges
1	5	1	6
0	10	84	94
Marges	15	85	100

Considérons les données ci-contre dans lesquelles :

Card $(A \cap \bar{B}) = n_{a\wedge\bar{b}} = 1$

Alors $q(a,\bar{b}) = -1{,}816$

et $\varphi(a,\bar{b}) = \frac{1}{\sqrt{2\pi}} \int_{-1{,}816}^{\infty} e^{-t^2/2}\, dt = 0{,}9653$

On dira dans ce cas que $(a \Rightarrow b)$ est admissible au niveau de confiance 96,5 %.
Remarquons deux valeurs particulières :

$\varphi(a,\bar{b}) \geq 0{,}95 \Leftrightarrow q(a,\bar{b}) \leq -1{,}65$

et $\varphi(a,\bar{b}) \geq 0{,}5 \Leftrightarrow q(a,\bar{b}) \leq 0.$

On notera que l'approche ci-dessus peut s'exprimer en termes de test d'hypothèse. Cependant, nous ne retenons pas cette voie qui, du fait de sa visée de prise de décision, limiterait les considérations qui vont suivre. Mais cette possibilité marque bien la différence avec l'approche de J. LOEVINGER (LOEVINGER [1947]) qui définissait la quasi-implication de a sur b par l'indice qui prend ses valeurs sur]-∞,1] :

$H(a,b) = 1 - \frac{nn_{a\wedge\bar{b}}}{n_a n_{\bar{b}}}$. Si $H(a,b)$ est "*proche*" de 1, l'implication est "*presque*" satisfaite.

Mais, comme on le voit, cet indice présente l'inconvénient, ne se référant pas à une échelle de probabilité, de ne pas fournir de seuil de vraisemblance et d'être invariant dans toute dilatation de E, A, B et $A \cap \bar{B}$. On a d'ailleurs la relation suivante entre $H(a,b)$ et $q(a,\bar{b})$:

$$\frac{q(a,\bar{b})}{H(a,b)} = -\sqrt{\frac{n_a n_{\bar{b}}}{n}} .$$

Cette limitation apparaît dans l'approche de J. PEARL (PEARL [1988]), de S. ACID et als (ACID [1991]) et A. GAMMERMAN, Z. LUO (GAMMERMAN A et LUO Z. [1991]). Chez ces derniers chercheurs, c'est l'écart entre la distribution conjointe entre a et b (et non a et $\bar{b}$) et la distribution produit qui tient lieu de critère comparatif.

Dans la recherche sur l'apprentissage de bases de connaissances de J.G. GANASCIA (GANASCIA J.G. [1991]), où "*l'incertitude*" sur l'implication $a \Rightarrow b$ est évaluée par l'indice : 2 Pr[$b \mid a$] - 1 et étendue à des classes de variables, la simplicité de l'approche de la quasi-implication se paie selon ces deux mêmes inconvénients. De plus, cet indice ne sépare pas, numériquement, deux implications dont l'une serait triviale et l'autre hautement informative.

1.2. Quelques propriétés du modèle de l'implication statistique.

A. LARHER dans sa thèse [1991] étudie et démontre différentes propriétés de q et φ. Les plus importantes, auxquelles nous adjoignons les propositions 5 et 6, sont les suivantes :

PROPOSITION 2. Si, n_a étant fixé et A inclus dans B, n_b tend vers n (B croît vers E), alors $\varphi(a,\bar{b})$ tend vers 0,5. Un prolongement par continuité nous permet donc de définir : si $B = E$ alors $\varphi(a,\bar{b}) = 0{,}5$.

PROPOSITION 3. *Pour toute variable a* :

$$0{,}95 \leq \varphi(a,\bar{a}) \leq 1 \Leftrightarrow n_a \in \left[\frac{n - \sqrt{n(n-11)}}{2}\right] ; \left[\frac{n + \sqrt{n(n-11)}}{2}\right] \qquad (1)$$

Or, pour n assez grand, l'intervalle ci-dessus est voisin de $[0,n]$, ce qui permet d'affirmer que l'implication statistique $a \Rightarrow a$ a un sens, tout en réservant une limite de confiance à la stabilité du caractère reproductible de la variable a.

PROPOSITION 4. *La relation* $\mathcal{R}$ *sur* V^2 *définie par* :

$\forall (a,b) \in V^2$ $\quad a \mathcal{R} b$ *dès que* $\varphi(a,\bar{b}) \geq 0{,}95$ *et que* n_a *vérifie* (1)

est réflexive, mais ni symétrique, ni antisymétrique, ni transitive.

Pour lui associer un graphe valué, sans cycle et transitif, on en prendra la restriction $\mathcal{R}'$ aux variables vérifiant la condition : si $a \mathcal{R}' b$ et $b \mathcal{R}' c$, alors l'arc (a,c) appartient au graphe seulement si $\varphi(a,\bar{c}) \geq 0{,}5$. $\mathcal{R}'$ définit alors un préordre partiel et permet une représentation claire de la relation d'implication statistique (cf. GRAS [1979]). On décrit dans [LERMAN I.C., GRAS R., ROSTAM H. 1981] l'algorithme qui permet de le construire. Ce seuil de 0,5 permet, en outre, la satisfaction d'un objectif d'accroissement informationnel. En effet, si I est l'incertitude associée aux variables a et b, il permet de vérifier $I(a \mid b) \leq I(a)$.

Remarque 1. Notre approche diffère également de celle de S. AMARGER et als (AMARGER S., DUBOIS D. et PRADE H. [1991]) qui, à partir d'une certaine inférence (une probabilité conditionnelle telle que $p(b \mid a)$ appartient à un intervalle, sans être parfaitement connue), induisent transitivement, de proche en proche, des probabilités conditionnelles sur un graphe incomplet, et cela sans la contrainte d'un seuil. Notre problématique, à l'opposé, vise l'analyse d'un tableau donné, sans ambition inductive a priori, mais en imposant un seuil qui autorise la fermeture transitive du graphe implicatif.

PROPOSITION 5. *Comparons le coefficient de corrélation $\rho(a,b)$ et l'indice $q(a,\bar{b})$*

$$\left| \textit{Supposons } q(a,\bar{b}) \neq 0. \textit{ Alors } \frac{\rho(a,b)}{q(a,\bar{b})} = -\sqrt{\frac{n}{n_b n_{\bar{a}}}} \right.$$

$$\textit{En effet, } q(a,\bar{b}) = \frac{n_{a\wedge\bar{b}} - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}} = \frac{n_a n_b - n n_{a\wedge b}}{\sqrt{n n_a n_{\bar{b}}}}$$

et $\rho(a,b) = \dfrac{n \cdot n_{a\wedge b} - n_a n_b}{\sqrt{n_a n_b\, n_{\bar{a}} n_{\bar{b}}}}$ d'où la relation entre $\rho(a,b)$ et $q(a,\bar{b})$

Ainsi $q(a,\bar{b}) = 0 \Leftrightarrow \rho(a,b) = 0$ *et* $\rho(a,b) \geq 0 \Leftrightarrow \varphi(a,\bar{b}) \geq 0{,}5$. Ceci signifie que implication et corrélation linéaire vont plutôt "*dans le même sens*". Cependant, on peut observer une croissance de l'implication en même temps qu'une décroissance de la corrélation. Ce qui montre bien, qu'outre la dépendance aux effectifs n, $n_{\bar{a}}$ et n_b, le rapport $\frac{\rho}{q}$ indique la non-coïncidence des deux concepts.

La double situation suivante l'illustre :

b_1 / a_1	1	0	Marges
1	96	4	100
0	56	44	100
Marges	152	48	200

$\rho_1(a_1,b_1) = 0{,}468$
$q_1(a_1,\bar{b}) = -4{,}082$

b_2 / a_2	1	0	Marges
1	94	6	100
0	52	48	100
Marges	146	54	200

$\rho_2(a_2,b_2) = 0{,}473$
$q_2(a_2,\bar{b}_2) = -4{,}041$

Par suite :
- d'une part, a_1 et b_1 sont moins corrélées que a_2 et b_2 ,
- d'autre part, l'intensité d'implication de a_1 sur b_1 est plus forte que celle de a_2 sur b_2 puisque $q_1 < q_2$.

PROPOSITION 6
Considérons le χ^2 d'indépendance des variables a et b ; alors

$$\frac{\chi^2}{q^2(a,\bar{b})} = \frac{n^2}{n_b\, n_{\bar{a}}} .$$

La démonstration est aisée, en particulier si l'on remarque que $\chi^2 = n\,\rho^2$.

On constate ainsi que les deux concepts χ^2 et q^2 ne se superposent pas. $q(a,\bar{b})$ est la racine carrée de la contribution à χ^2 de la case $(a,\bar{b})$ du tableau de croisement des 2 variables a et b. La seule considération, non relative, de χ^2 et des effectifs des 4 cases de ce tableau, comme le font souvent les psychologues, ne peut donc pas rendre compte précisément de l'implication.

Remarque 2. Etudions la sensibilité de q aux faibles variations d'effectif. Supposons pour cela que, par exemple, $n_{a\wedge\bar{b}}$ devienne $n'_{a\wedge\bar{b}} = n_{a\wedge\bar{b}} + k$ où $k \in \mathbb{Z}$ sans que varient n_a et $n_{\bar{b}}$. Alors :

$$q'(a,\bar{b}) = q(a,\bar{b}) + \frac{k}{\sqrt{\frac{n_a\, n_{\bar{b}}}{n}}}.$$

L'effet de l'erreur de mesure k est atténué en $\sqrt{\frac{n}{n_a\, n_{\bar{b}}}}$, ce qui permet dans la plupart des cas de maintenir la validité d'une implication au même seuil ou à un seuil voisin.

1.3. Extension à des variables numériques

Dans les paragraphes précédents, nous avons étudié le croisement entre un ensemble de variables et un ensemble de sujets, croisement qui, à la variable binaire a, associe la valeur $a(x) = 1$ (x satisfait a) ou 0 (x ne satisfait pas a).

Dans sa thèse, A. LARHER considère deux autres types de variables :
- *variables modales* associées à une logique multivalente où les valeurs de vérité ne sont plus seulement "*vrai* (1)" ou "*faux* (0)" mais toute valeur de [0,1]. La valeur prise par la variable a sur le sujet x est alors $a(x)$ élément de [0,1]. On retrouve ici un type de variable que E. DIDAY (DIDAY E. [1991]) axiomatise et étudie en profondeur dans le cadre de l'analyse des connaissances (au sens de l'I.A.). Mais, notre théorisation sera plus élémentaire ;

- *variables quantitatives* spécifiant un certain nombre de fois où, par exemple, la variable a est présente chez le sujet x. Par exemple, a étant une procédure de résolution, $a(x)$ est le nombre de fois où x l'emploie au cours de la résolution d'un problème.

En fait, une suggestion de I.C. LERMAN confortant notre propre approche, permet d'interpréter ces variables en terme de *variables fréquentielles* définies de la façon suivante :

Soit (a,b) $\in V^2$ et i $\in$ E, α (resp. β) la valeur maximum de a (resp. b) sur E. α_i (resp.β_i) la *valeur observée* de a (resp. b) sur i. Posons :

$$\alpha'_i = \frac{\alpha_i}{\alpha}, \beta_i' = \frac{\beta_i}{\beta}, v_a = \sum_{i\in E} \alpha'_i \ , \ v_b = \sum_{i\in E} \beta'_i \ , \ v_{a\wedge\bar{b}} = \sum_{i\in E} \alpha'_i\,(1-\beta'_i) \text{ et } v_{\bar{b}} = n - v_b .$$

Ces valeurs pour $\alpha = \beta = 1$, α_i et $\beta_i = 1$ ou 0 étendent alors le cas binaire où les effectifs fréquentiels sont n_a, n_b et $n_{a\wedge\bar{b}}$. En effet, si 1_A et 1_B sont les fonctions indicatrices des sous-ensembles d'individus possèdant respectivement les caractères a et b, l'effectif $n_{a\wedge\bar{b}}$ s'écrit :

$$n_{a\wedge\bar{b}} = \sum_{i\in E} 1_A(i)\ [1 - 1_B(i)]$$

Donc posons, comme pour les variables binaires :

$$q(a,\bar{b}) = \frac{v_{a\wedge\bar{b}} - \frac{v_a\, v_{\bar{b}}}{n}}{\sqrt{\frac{v_a\, v_{\bar{b}}}{n}}}$$

qui sera, pour $v_a \le v_b$, l'indicateur retenu pour l'implication statistique de a sur b.

Nous en donnerons un exemple dans le paragraphe 3.3.

2. IMPLICATION ENTRE CLASSES DE VARIABLES.

Elle ne prend véritablement son sens qu'à condition qu'à l'intérieur de chaque classe de variables dont on examine la relation avec d'autres, existe une certaine "cohésion" entre les variables qui la constituent, ceci afin que le "flux" implicatif d'une classe $\mathfrak{A}$ sur une classe $\mathfrak{B}$ soit nourri d'un "flux" interne à $\mathfrak{A}$ et alimente un "*flux*" interne à $\mathfrak{B}$. Cette cohésion, généralement nourrie de cohérence sémantique ou, dans le cas de la didactique, de conditions psychologiques, cognitives, situationnelles, etc., doit se traduire ici par une mesure (quantitative). On pourrait penser qu'un ensemble d'indices de similarité assez élevés entre les éléments de la classe serait un bon indicateur de cohésion. Nous ne retenons pas cette approche qui ne rendrait compte que d'une cohésion de profils symétriquement comparables, ne restituant pas une dynamique interne orientée (donc non symétrique). Or, nous disposons avec les intensités d'implication entre variables d'un instrument de mesure d'un emboîtement de deux parties d'une population E. Par exemple, si dans la classe à 3 éléments a, b, c, on observe : $\varphi(a,\bar{b}) = 0{,}97$, $\varphi(b,\bar{c}) = 0{,}95$, $\varphi(a,\bar{c}) = 0{,}92$, , on pourra dire que la classe orientée de a vers c admet une bonne cohésion ; ce qui ne serait pas le cas si $\varphi(a,\bar{b}) = 0{,}82$, $\varphi(b,\bar{c}) = 0{,}38 < \varphi(b,\bar{c}) = 0{,}70$ et $\varphi(a,\bar{c}) = 0{,}48 > \varphi(c,\bar{a})$. C'est donc cette voie que nous choisissons pour une cohésion implicative donc orientée, comme peut l'être une filiation procédurale ou une genèse. Nous verrons ensuite quel indicateur permettrait de rendre compte d'une extension aux classes, de la notion d'implication.

2.1. Cohésion implicative

Afin d'en améliorer l'intuition, nous la définirons progressivement pour 2, puis 3 et r éléments de classe.

La cohésion, se voulant indicateur d'ordre implicatif au sein d'une classe de variables, s'oppose en cela au"désordre" dont rend compte l'entropie d'une expérience aléatoire. Rappelons au sujet de celle-ci que, X étant une variable aléatoire prenant ses valeurs dans $S = \{m_1,m_2,..,m_k\}$ muni de la loi $\{p_1,p_2,...,p_k\}$, l'entropie est l'espérance mathématique de la variable $I(X)$ prenant les valeurs $I(m1), I(m2),...,I(m_k)$; $I(m_j)$ est l'incertitude sur $\{m_j\}$ ou information apportée par la réalisation de $\{m_j\}$. Ainsi :

$$\mathcal{E}I[(X)] = \sum_{j=1}^{k} - p_j \log_2 p_j$$

est **l'entropie de l'expérience**.

2.1.1. Cas de 2 éléments : classe (a,b).

Supposons $n_a \leq n_b$. Nous allons définir la cohésion du couple (a,b).

Soit χ la variable aléatoire indicatrice de l'événement $[Q(a,\bar{b}) \geq q(a,\bar{b})]$. Alors :

$$\Pr(\chi = 1) = \varphi(a,\bar{b}) = p$$

$$\text{et } \Pr(\chi = 0) = 1 - \varphi(a,\bar{b}) = 1 - p$$

L'entropie ou incertitude de cette expérience est alors :

$$\mathcal{E}[I(\chi)] = - p \log_2 p - (1 - p) \log_2(1 - p)$$

Par exemple :

. si $\varphi(a,\bar{b}) = p = 0{,}95$, alors :

$$\mathfrak{E}[I(X)] = \frac{-0{,}95 ln\ 0{,}95 - 0{,}05 ln\ 0{,}05}{ln\ 2} = 0{,}286$$

. si $\varphi(a,\bar{b}) = 1$, alors $\mathfrak{E}[I(\chi)] = 0$ en convenant que $0\ ln\ 0 = 0$

. si $\varphi(a,\bar{b}) = 0{,}5$, alors $\mathfrak{E}[I(\chi)] = 1$ (entropie maximale).

Mais si $\varphi'(a,\bar{b}) = 1 - \varphi(a,\bar{b})$, alors $\mathfrak{E}[I(\chi)] = \mathfrak{E}[I(1-\chi)]$.

Précisément, en posant : $\mathfrak{E} = f(p) = -p\ \log_2 p - (1-p)\log_2(1-p)$, on a : $f(1-p) = f(p)$.

L'entropie est donc symétrique par rapport à p = 0,5 .

De plus, $\frac{df}{dp} = \log_2 \frac{1-p}{p}$ pour $p \in]0,1[$ donc $\mathfrak{E}$ croît de 0 à 1 sur]0;0,5] et décroît de 1 à 0 sur [0,5;1[.

Aussi, la propriété de symétrie de $\mathfrak{E}$ allant à l'encontre de la dissymétrie de la quasi-implication, nous retiendrons finalement, comme indicateur de cohésion, l'application c définie sur $V \times V$ par :

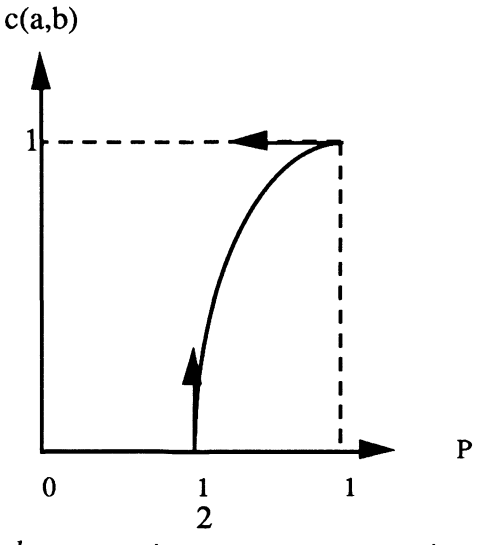
DÉFINITION 2. *La cohésion du couple de variables (a,b) tel que $n_a \leq n_b$ est le nombre c(a,b) où :*

. si $\varphi(a,\bar{b}) = p \geq 0{,}5$, $c(a,b) = [1 - [p\ log_2 p + (1-p)\ log_2(1-p)]^2]^{1/2} = \sqrt{1 - \mathfrak{E}^2}$
. et si $\varphi(a,\bar{b}) = p < 0{,}5$, $c(a,b) = 0$ (*absence de cohésion*).

La fonction "*carré de l'entropie*" est choisie pour des raisons de renforcement du contraste sur [0,1]. Nous prenons la racine carrée de son complément à 1 pour donner à la cohésion la dimension de l'entropie et pour accroître sa valeur numérique (en effet pour $x \in [0,1]$, $\sqrt{1-x^2} \geq 1 - x$)

2.1.1.1.*Etude aux limites*

Posons : c(a,b) = g($\mathfrak{E}$) = g[f(p)]



$\frac{dg}{d\mathcal{E}} = -\frac{\mathcal{E}}{\sqrt{1-\mathcal{E}^2}}$ *et* $\frac{dc}{dp} = \frac{-\mathcal{E}}{\sqrt{1-\mathcal{E}^2}} \times log_2 \frac{1-p}{p}$.

$c(a,b)$ croît donc de 0 à 1 quand $p = \varphi(a,\bar{b})$ croît de 0,5 à 1. La fonction c de p est continue en $p = \frac{1}{2}$

Nous prolongeons par continuité en prenant :
$c(a,b) = 1$ lorsque $p = 1$ (c'est-à-dire lorsque l'implication $a \Rightarrow b$ est stricte).

Remarque 3. Rappelons un résultat démontré dans la thèse d'A. LARHER :
si $n_a \leq n_b$ et si $q(a,\bar{b}) \leq 0$, alors l'intensité d'implication de a sur b est supérieure à celle de b sur a. Ce qui signifie que :

$$n_a \leq n_b \text{ et } q(a,\bar{b}) \leq 0 \Rightarrow \varphi(a,\bar{b}) = \max(\varphi(a,\bar{b}), \varphi(b,\bar{a})).$$

Appelant **classe** (a,b) le couple (a,b) tel que $n_a \leq n_b$, la cohésion de la classe (a,b) est définie sans équivoque à partir de la plus grande des intensités relatives à $(a \Rightarrow b)$ et $(b \Rightarrow a)$.

D'où la :

DÉFINITION 2'

La cohésion de la classe (a,b) est le nombre $c(a,b)$ tel que :

. *si* $p = max[\varphi(a,\bar{b}), \varphi(b,\bar{a})] \geq 0,5$ *et* $\mathcal{E} = -p\, log_2 p - (1-p)\, log_2(1-p)$
$c(a,b) = \sqrt{1-\mathcal{E}^2}$
. *si* $p = 1$ $c(a,b) = 1$
. *si* $p \leq 0,5$ $c(a,b) = 0$

Remarque 4. Lorsque $b = a$ et que $n_a \in \left[\frac{n - \sqrt{n(n-11)}}{2} ; \frac{n + \sqrt{n(n-11}}{2}\right]$ nous avons vu que (cf. p. 11) $0,95 \leq \varphi(a,a) \leq 1$. La cohésion de la classe (a,a) a donc un sens et, en général, puisque $\lim_{p \to 1^-} c = 1$, cette cohésion est très voisine de 1. Par définition, nous poserons donc :

$$\forall a \in V \qquad c(a,a) = 1.$$

2.1.2. Cas de 3 éléments a, b et c.

Six valeurs d'intensité correspondent a priori à l'ensemble $\dot{A} = \{a,b,c\}$:

$$\varphi(a,\bar{b}),\varphi(a,\bar{c}),\varphi(b,\bar{a}),\varphi(c,\bar{a}) \text{ et } \varphi(c,\bar{b})$$

Nous souhaitons que l'indice de cohésion implicative contienne l'information révélée par les relations implicatives binaires entre *tous* les éléments de l'ensemble $\dot{A}$. Mais, en même temps, pour conserver la dynamique dissymétrique de l'implication, seule la relation la plus puissante entre deux éléments quelconques reste pertinente par rapport à notre objectif. Par suite, parmi toutes les associations 3 à 3 ne faisant intervenir qu'une fois chaque couple d'éléments de $\{a,b,c\}$ et restituant au mieux la puissance de certaines implications, nous retenons les valeurs :

$$\max\left[\varphi(a,\bar{b}),\varphi(b,\bar{a})\right], \max\left[\varphi(a,\bar{c}),\varphi(c,\bar{a})\right] \text{ et } \max\left[\varphi(b,\bar{c}),\varphi(c,\bar{b})\right]$$

Comme précédemment, dans le cas où ils sont supérieurs ou égaux à 0,5, les maxima obtenus sont compatibles avec l'ordre des effectifs n_a,n_b et n_c. Par exemple, si $n_a \leq n_b \leq n_c$, les trois maxima sont : $\varphi(a,\bar{b}),\varphi(a,\bar{c})$ et $\varphi(b,\bar{c})$.

DÉFINITION 3. *Le couple $\mathcal{C} = ((a,b),c)$ sera encore appelé* classe *et sa cohésion implicative sera définie ainsi* :

$$C(\mathcal{C}) = [c(a,b) \times c(b,c) \times c(a,c)]^{1/3}$$

moyenne géométrique des cohésions des classes à 2 éléments

La préférence accordée à la moyenne géométrique plutôt qu'à la moyenne arithmétique tient à notre volonté, d'une part d'obtenir une cohésion nulle pour une classe dès que la cohésion d'un de ses couples est nulle, c'est-à-dire dès que les implications mutuelles sont inférieures ou égales à 0,5, d'autre part de "*ramener*" $C(\mathcal{C})$ au voisinage de 1 lorsque les cohésions des couples sont assez fortes.

2.1.3. Cas de r éléments a_1, a_2,...,a_r.

Nous opérons comme ci-dessus, c'est-à-dire en retenant les maxima des intensités d'implication entre 2 éléments quelconques de l'ensemble $\dot{A} = \{a_1,a_2,...,a_r\}$. A ces maxima sont associés les cohésions implicatives des couples et l'ordre induit sur $\dot{A}$ par les effectifs n_{a_1}, n_{a_2}, ..., n_{a_r}. Par exemple, si $n_{a_1} \leq n_{a_2} \leq ... \leq n_{a_r}$, nous appellerons *classe* le couple $\mathcal{C} = (((a_1,a_2),a_3),...,a_r)$, et comme il y a $\frac{r(r-1)}{2}$ paires, sa cohésion implicative sera :

DÉFINITION 4

$$C(\mathcal{C}) = \left[\prod_{\substack{i\in\{1,...,r-1\} \\ j\in\{2,...,r\} \\ j>i}} c(a_i,a_j)\right]^{\frac{2}{r(r-1)}}$$ moyenne géométrique des cohésions de classes à 2 éléments

2.2. Implication entre classes.

Nous souhaitons que l'implication entre deux classes se constitue à partir des informations suivantes :
- les cohésions respectives des 2 classes,
- une intensité d'implication extrémale des éléments d'une classe sur les éléments de l'autre,
- les cardinaux respectifs des 2 classes.

Chacune de ces informations crédite l'indice que nous retiendrons si :
- l'indice croît avec les cohésions de chaque classe et s'annule lorsque la cohésion de l'une d'entre elles est nulle,
- l'indice croît avec la liaison extrémale (minimale si l'on vise un degré d'exigence élevé, maximale si l'on recherche une souplesse réaliste),
- l'indice décroît avec les cardinaux des classes, eu égard à la prise en compte d'une liaison maximale.

Posons : . $\dot{A}$ et $\dot{B}$ deux parties disjointes : $\dot{A} = \{a_1,...,a_r\}$ et $\dot{B} = \{b_1,...,b_s\}$,
. $\mathcal{A}$ et $\mathcal{B}$ les classes qui leur sont respectivement associées ,
. $C(\mathcal{A})$ et $C(\mathcal{B})$ leurs cohésions respectives.

Conformément aux lois de probabilité des sup. de variables aléatoires, a priori uniformément distribuées, nous définissons l'*indice d'implication* $\psi(\mathcal{A},\mathcal{B})$ *de la classe* $\mathcal{A}$ *vers la classe* $\mathcal{B}$ *par :*

DÉFINITION 5.

$$\psi(\mathcal{A},\mathcal{B}) = \left\{ \sup_{\substack{i\in\{1,...,r\} \\ j\in\{1,...,s\}}} \varphi(a_i,\bar{b}_j) \right\}^{rs} \times \left[C(\mathcal{A}) \times C(\mathcal{B}) \right]^{\frac{1}{2}}$$

L'expression $[C(\mathcal{A})\ C(\mathcal{B})]^{\frac{1}{2}}$ représente la cohésion moyenne (géométrique) de $\mathcal{A}$ et $\mathcal{B}$; elle intègre les informations de cohésivité des 2 classes en jeu ; de plus, cette expression est telle que si $C(\mathcal{A})$ et $C(\mathcal{B})$ sont simultanément multipliées par k, alors $\psi(a,b)$ est multipliée par k.

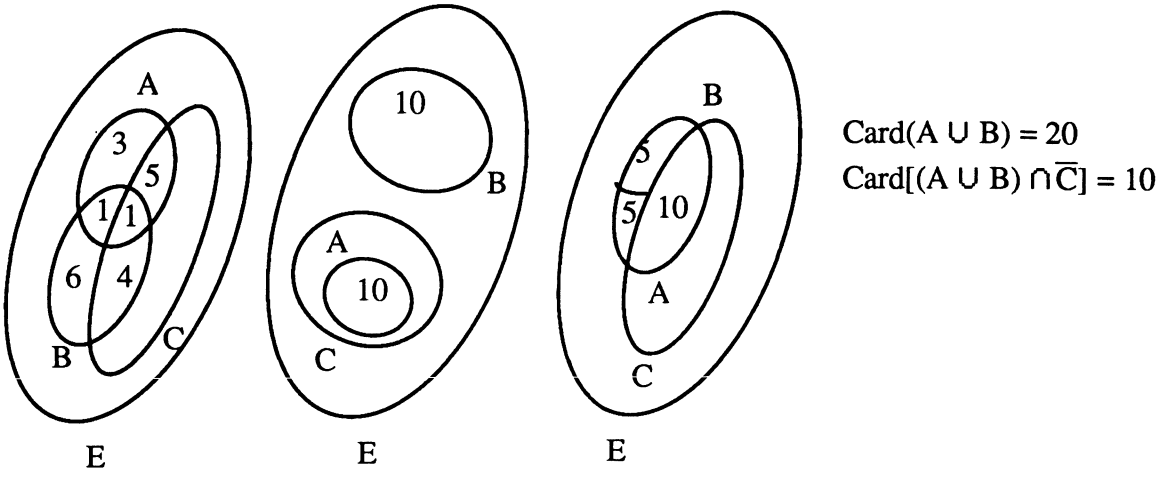
REMARQUES

1°) Si $\dot{A}$ et $\dot{B}$ sont réduits à des singletons, et donc $\mathcal{A}$ et $\mathcal{B}$ à des couples d'éléments égaux, cette formule vérifie la définition de l'implication entre variables puisque dans ce cas $C(\mathcal{A}) = C(\mathcal{B}) = 1$; l'indice choisi coïncide alors avec l'intensité d'implication, soit : $\psi((a,a),(b,b)) = \varphi(a,\bar{b})$.

2°) La méthode de construction de l'indice d'implication ci-dessus montre à l'évidence qu'elle est indépendante de la nature des variables traitées, que celles-ci soient binaires, non binaires ou fréquentielles.

3°) Un critère d'implication entre classes basé, dans le cas des variables binaires (ou attributs), sur la réunion ensembliste des ensembles d'individus possédant ces attributs n'est pas pertinent. D'une part, il privilégie les seules variables de cette nature, d'autre part il ne rend pas compte de la cohésion des classes.

Par exemple, si a, b et c sont des attributs représentés respectivement par les parties A, B et C ci-dessous et si, de plus, Card($A \cup B$), Card C, et Card[$(A \cup B) \cap \overline{C}$] restent constants, on peut obtenir des situations très différentes s'opposant à l'idée intuitive de l'implication entre classes telle que nous la concevons. La cohésion de (a,b) et les positions respectives de A, B et C dans 3 cas différents montrent de façon évidente l'inadéquation du choix de la réunion pour signifier l'implication :



La cohésion de la classe (a,b) et la relation implicative de (a,b) sur c sont manifestement très différentes dans les 3 cas, alors que ceux-ci sont très comparables selon le critère de la réunion ensembliste.

4°) Les moyens de calcul à développer pour examiner tous les indices d'implication de classes dans un ensemble de x variables (définies sur une population de n individus) croissent exponentiellement avec x ; en effet, l'ensemble des parties à considérer contient 2^x éléments. Aussi, dans la réalité, nous pensons raisonnable de procéder de la façon suivante :
- calculer les implications entre variables,
- construire puis analyser le graphe d'implication (cf. p. 11),
- émettre des hypothèses quant à la cohérence de formation de classes par rapport à la problématique initiale,
- constituer de telles classes de variables (items ou modalités de réponse dans le cas d'un questionnaire) et évaluer les valeurs des indices d'implication entre ces classes prises 2 à 2 ; éventuellement, modifier sensiblement les contenus de ces classes pour améliorer les valeurs des indices d'implication. Un indicateur statistique, qui reste à déterminer, pourrait servir à définir un test d'arrêt de modification.

Par exemple, dans le cas où les variables seraient des procédures (disjointes ou non, mais identifiables en présence-absence chez chacun des individus), on pourrait obtenir des relations entre classes de procédures, gommant les effets microscopiques de procédures trop parcellisées et rendant compte de démarches générales, voire de conceptions, ayant une certaine stabilité.

Notons à ce sujet que la condition de disjonction de $\dot{A}$ et $\dot{B}$ exprimée plus haut peut être levée sans affecter radicalement les objectifs déclarés à propos de l'implication entre classes : en effet, la cohésion de classes est un garant d'homogénéité "*dirigée*" qui relativise l'incidence qu'aurait le premier facteur du produit définissant l'implication $\psi(a,b)$ où $\mathfrak{A}$ et $\mathfrak{B}$ sont les classes respectivement associées aux ensembles $\dot{A}$ et $\dot{B}$.

Pour terminer ce paragraphe, soulignons l'intérêt des notions de cohésion et d'implication entre classes pour l'analyse implicative.

Soit une classe $\mathcal{A}$ de variables de cohésion non nulle. Si l'une de ses variables, a, admet une implication inférieure à 0,5 sur une variable d'une classe $\mathcal{B}$, le graphe implicatif ne pourra rendre compte de cette implication en raison de l'exigence d'une intensité supérieure ou égale à 0,5 pour la fermeture transitive du graphe (cf. p. 11). Par contre, l'implication entre classes ne sollicitant que le sup des implications dirigées de a vers b, l'attribut a va apparaître dans l'implication $\mathcal{A} \Rightarrow \mathcal{B}$. Ainsi, nous conservons une information qui aurait disparu si l'on s'était contenté d'envisager l'implication de variables seules.

2.3. Etude de la vraisemblance de $\psi(\mathcal{A}, \mathcal{B})$.

Dans ce paragraphe, nous cherchons des critères permettant, à partir d'une valeur observée dans une situation, toujours dans l'hypothèse d'une absence de lien a priori, d'accorder un sens implicatif à la liaison non symétrique de l'ensemble de variables $\dot{A}$ (classe associée : $\mathcal{A}$) sur l'ensemble de variables $\dot{\mathcal{B}}$ (classe associée : $\mathcal{B}$). En fait, à travers le questionnement de la valeur de $\psi(\mathcal{A}, \mathcal{B}) = \left[\sup_{i,j} \varphi(a_i, \bar{b}_j)\right]^{rs} \times \left[C(\mathcal{A})\, C(\mathcal{B})\right]^{\frac{1}{2}}$, nous avons la même attitude que celle que nous avons eue à l'égard de $\varphi(a,\bar{b}) = 1 - Pr\left[Q(a,\bar{b}) \leq q(a,\bar{b})\right]$. Pour cela, il nous faut examiner les lois régissant les variations des variables en jeu : intensités d'implication et cohésions implicatives. La complexité des calculs nous contraint à des études limitées mais déjà significatives de l'implication de $\mathcal{A}$ sur $\mathcal{B}$ ou "*degré d'étonnement*" des valeurs observées (cf. p. 20).

2.3.1. Loi de variation de $c(a,b)$.

Nous menons cette étude dans le cas des variables binaires, mais cette étude serait très proche dans les autres cas.

Lorsque la cohésion d'une classe $\mathcal{A}$ est faible, la liaison implicative entre les éléments est cependant loin d'être négligeable. En effet, cette valeur faible reste cependant le témoin qu'il existe toujours entre 2 éléments quelconques de l'ensemble associé $\dot{A}$ une implication unidirectionnelle au moins égale à 0,5. Précisons cela dans le cas où $\dot{A} = \{a,b\}$.

PROPOSITION 7. *Posant* $\alpha' = 1 - \alpha \in [0,1]$, Φ et $\Gamma(a,b)$ *les variables aléatoires dont les réalisations empiriques respectives sont l'intensité d'implication* $\varphi(a,\bar{b})$ *et la cohésion* $c(a,b)$ *alors* :

$$Pr\left[\Gamma(a,b) > \alpha'\right] = Pr\left[\Phi^{\Phi}(1-\Phi)^{1-\Phi} > \frac{1}{2^{\sqrt{1-\alpha'^2}}}\right]$$

a) Deux attributs a et b étant donnés, considérons en effet la variable aléatoire Φ dont la valeur observée, $n_{a \wedge \bar{b}}$ étant connu, est $\varphi(a,\bar{b})$ Ainsi, $[\Phi \geq 1 - \alpha]$ si et seulement si $[Q(a,\bar{b} \leq \varphi'^{-1}(\alpha)]$ où φ'^{-1} est la fonction réciproque de l'intégrale gaussienne :

$$\alpha = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{q(a,\bar{b})} e^{-t^2/2}\, dt$$

Revenant à la nature "*poissonnienne*" de la variable $\mathrm{Card}(X \cap \overline{Y})$ définie au § 1.1, représentant le cardinal aléatoire de $A \cap \overline{B}$ de valeur observée $n_{a\wedge\bar{b}}$, on obtient :

$$\Pr[\Phi \geq 1 - \alpha] = \Pr[\mathrm{Card}(X \cap \overline{Y}) \leq k] .$$

$$\text{Soit } \Pr[\Phi \geq 1 - \alpha] = \sum_{l=0}^{k} e^{-\lambda} \frac{\lambda^l}{l!} \text{ avec } \lambda = \frac{n_a n_{\bar{b}}}{n}$$

Bien entendu, pour n assez grand, nous retrouvons : $\Pr[\Phi \geq 1 - \alpha] \simeq \alpha$ donc Φ est uniformément distribuée sur [0,1].

Par exemple, si $n = 100$, $n_a = 15$, $n_b = 20$:
$\Pr[\Phi \geq 0{,}9] = 0{,}086$ $(k < 7{,}57)$
$\Pr[\Phi \geq 0{,}95] = 0{,}043$ $(k < 6{,}28)$.

b) Soit $\alpha' = 1 - \alpha \geq 0$. Alors, revenant sur les variations de $c(a,b)$ en fonction de $\varphi(a,\bar{b})$, examinons la loi de la variable aléatoire $\Gamma(a,b)$, de réalisation $c(a,b)$, en fonction de celle de Φ .

$$\Pr[\Gamma(a,b) > \alpha'] = \Pr\left\{1 - [\Phi \log_2 \Phi + (1 - \Phi)\log_2(1 - \Phi)]^2 > \alpha'^2 \text{ et } \Phi \geq 0{,}5\right\}$$

$$= \Pr\left\{1 - \frac{1}{(\ell n\, 2)^2}[\Phi\, \ell n\, \Phi + (1 - \Phi)\ell n(1 - \Phi)]^2 > \alpha'^2 \text{ et } \Phi \geq 0{,}5\right\}$$

$$= \Pr\left\{(1 - \alpha'^2)(\ell n\ 2)^2 > [\Phi\, \ell n\ \Phi + (1 - \Phi)\ell n\,(1 - \Phi)]^2 \text{ et } \Phi \geq 0{,}5\right\}$$

Supposant $\Phi \geq 0{,}5$ réalisé :

$$\Pr[\Gamma(a,b) > \alpha'] = \Pr\left\{\underbrace{\Phi\, \ell n\, \Phi + (1 - \Phi)\ell n\ (1 - \Phi)}_{\leq 0} < \ell n\ 2^{\sqrt{1 - \alpha'^2}}\right\} .$$

$$= \Pr\left\{\Phi\ \ell n\ \Phi + (1 - \Phi)\ \ell n\ (1 - \Phi) > -\ \ell n\ 2^{\sqrt{1 - \alpha'^2}}\right\}$$

Mais, $\Phi\ \ell n\ \ \Phi + (1 - \Phi)\ \ell n\ (1 - \Phi) = \ell n\ \Phi^{\Phi}(1 - \Phi)^{1-\Phi}$

d'où $$\Pr[\Gamma(a,b) > \alpha'] = \Pr\left[\Phi^{\Phi}(1 - \Phi)^{1-\Phi} > \frac{1}{2^{\sqrt{1 - \alpha'^2}}}\right]$$

Exemples.

* si $\alpha' = 0{,}95$, $\frac{1}{2^{\sqrt{1 - \alpha'^2}}} = 0{,}8054$

D'où $\Pr[\Gamma(a,b) > 0{,}95] = \Pr[\Phi > 0{,}945] = 0{,}055$.

* Réciproquement, $\Pr[\Phi > 0{,}95] = \Pr[\Phi^{\Phi}(1 - \Phi)^{1-\Phi} > 0{,}820]$
$= \Pr[\Gamma(a,b) > 0{,}958]$

en effet $\frac{1}{2^{\sqrt{1-\alpha'^2}}} > 0{,}820$

pour $2^{\sqrt{1-\alpha'^2}} < 1{,}2195$

soit $\sqrt{1-\alpha'^2} < 0{,}28629$

ou encore : $\alpha' > 0{,}958.$

* On trouve également :

$\Pr[\Gamma(a,b) > 0{,}9] = 0{,}09$

$\Pr[\Gamma(a,b) > 0{,}8] = 0{,}15.$

2.3.2. Etude de la cohésion $C(\mathcal{Q})$.

$\mathcal{Q}$ est la classe associée à l'ensemble $A = \{a_i\}_{i=1,\ldots,r}$; rappelons que, $c(a_i, a_j) = (1-\mathcal{E}^2)^{\frac{1}{2}}$ étant la meilleure des 2 cohésions $c(a_i,a_j)$ et $c(a_j,a_i)$ et $\mathcal{E}$ l'entropie associée à l'implication $a_i \Rightarrow$ aj, alors

soit encore :

$$C(\mathcal{Q}) = \left[\prod_{i,j} c(a_i,a_j)\right]^{\frac{2}{r(r-1)}}$$

$$\ell n\, C(\mathcal{Q}) = \frac{2}{r(r-1)} \sum_{i,j} \ell n\; c(a_i,a_j)\,.$$

Dans une hypothèse d'absence de lien, les variables aléatoires $\Gamma(a_i,a_j)$, dont les cohésions $c(a_i,a_j)$ sont des réalisations, sont indépendantes et $\ell n\ \Gamma(\mathcal{Q})$ a pour loi le produit de convolution des lois de ces variables $\Gamma(a_i,a_j)$, au coefficient $\frac{2}{r(r-1)}$ près. Or nous avons vu la difficulté d'accéder à la loi de $\Gamma(a_i,a_j)$. Il nous sera, par contre, loisible d'effectuer des simulations pour estimer la loi de $\Gamma(\mathcal{Q})$. Cette tâche est une composante de la thèse d'André TOTOHASINA, sous la direction de R.GRAS, à Rennes.

Notons cependant que :

si $\forall$ i, $\forall$ j, $\quad c(a_i,a_j) > 0{,}95, \quad$ alors $C(\mathcal{Q}) > 0{,}95.$

2.3.3. Etude du terme $\left[\sup_{i,j} \varphi(a_i,\bar{b}_j)\right]^{rs}$

Comme nous l'avons fait pour l'intensité d'implication entre attributs, nous considérons chaque probabilité $\varphi(a_i,\bar{b}_j)$ comme la réalisation d'une variable aléatoire $\Phi_{i,j}$ uniformément distribuée sur [0,1]. Dans l'hypothèse d'absence de lien a priori entre les attributs, les variables aléatoires $\Phi_{i,j}$ sont indépendantes. Par suite, si $S = \sup_{i,j} \Phi_{i,j}$,

alors $\Pr[S \geq 1 - \alpha] = \alpha$

2.3.4. Etude de $\psi(\mathcal{A},\mathcal{B})$, dans le cas où $\mathcal{A}$ et $\mathcal{B}$ ne contiennent que 2 éléments au plus.

Rappelons que : $\psi(\mathcal{A},\mathcal{B}) = \left[(\sup_{i,j} \varphi(a_i,\bar{b}_j)\right]^{rs} (C(\mathcal{A})\, C(\mathcal{B}))^{\frac{1}{2}}$

donc : $$\Pr[\Phi > \alpha'] \leq \Pr\left[(\sup_{i,j} \Phi_{i,j})^{rs} > \alpha' \text{ et } C(\mathcal{A}) > \alpha'^2 \text{ et } C(\mathcal{B}) > \alpha'^2\right]$$

$$\leq \Pr\left[(\sup_{i,j} \Phi_{i,j})^{rs} > \alpha'\right] . \Pr[C(\mathcal{A}) > \alpha'^2] . \Pr[C(\mathcal{B}) > \alpha'^2]$$

dans une hypothèse d'indépendance a priori.

Par exemple,

1°) si nous souhaitons un "*degré d'étonnement*" ou indice d'implication de classes $\psi(\mathcal{A},\mathcal{B})$ qui soit supérieur ou égal à 0,97, il suffit de prendre $\alpha' > 0{,}7$ car alors : $C(\mathcal{A})$ et $C(\mathcal{B})$ sont supérieures à 0,49 ; donc pour chaque classe $\mathcal{A}$ et $\mathcal{B}$ on a : $\varphi^{\varphi}(1-\varphi)^{1-\varphi} > 0{,}546$, soit $\varphi > 0{,}7$, où φ est l'une quelconque des réalisations de $\Phi_{i,j}$.

Par suite, $\Pr[\Gamma(\mathcal{A}) > 0{,}49] = \Pr[\Phi^{\Phi}(1-\Phi)^{1-\Phi} > 0{,}546] = \Pr[\Phi > 0{,}7] \leq 0{,}3$.

Dans ce cas, $\Pr\left[(\sup_{i,j} \Phi_{i,j})^{rs} > 0{,}7\right] \times \Pr[\Gamma(\mathcal{A}) > 0{,}49] \times \Pr[\Gamma(\mathcal{B}) > 0{,}49]$ est inférieure ou égale à $0{,}3 \times 0{,}3 \times 0{,}3 \leq 0{,}03$ et $\Pr[\Phi > \alpha] \leq 0{,}03$, soit un "*degré d'étonnement*" supérieur ou égal à 0,97.

Ceci signifie qu'une observation $\psi(\mathcal{A},\mathcal{B}) > 0{,}7$, pour $\mathcal{A}$ et $\mathcal{B}$ constituées de couples, constitue un événement d'intensité d'implication de classes supérieure à 0,97.

2°) Pour $\psi(\mathcal{A},\mathcal{B}) \geq 0{,}95$, il suffit de prendre $\alpha' \geq 0{,}63$ car alors : $C(\mathcal{A}) > 0{,}4$ et $C(\mathcal{B}) > 0{,}4$ et donc pour chaque classe $\mathcal{A}$ et $\mathcal{B}$, on a $\varphi^{\varphi}(1-\varphi)^{1-\varphi} > 0{,}53$ soit $\varphi > 0{,}65$ et $\Pr[\Gamma(\mathcal{A}) > 0{,}4] \leq 0{,}35$.

Dans ce cas : $\Pr[(\sup\Phi_{i,j})^{rs} > 0{,}63] \times [\Pr[\Gamma(\mathcal{A}) > 0{,}4] \times \Pr[\Gamma(\mathcal{B}) > 0{,}4]$

$\leq 0{,}37 \times 0{,}35 \times 0{,}35 \leq 0{,}05$.

Une observation $\psi(\mathcal{A},\mathcal{B}) \geq 0{,}63$ constitue un événement d'intensité d'implication de classes supérieure à 0,95.

Notons que cette intensité d'implication de la classe $\mathcal{A}$ sur la classe $\mathcal{B}$ joue par rapport à la valeur observée, indice d'implication $\psi(\mathcal{A},\mathcal{B})$, le même rôle que celui joué par l'intensité $\varphi(a,\bar{b})$, dans le cas de 2 attributs, par rapport à l'indice $q(a,\bar{b})$. Cette intensité mesure la qualité de l'"*étonnement*" que nous avons face à une observation dans l'hypothèse d'absence de lien.

2.4. Agrégations successives des classes.

2.4.1. Algorithme.

Selon l'objectif classique des méthodes hiérarchiques, nous allons définir un algorithme d'agrégations successives des classes.

1ère étape : considérant toutes les paires de variables, on détermine le couple (a_i,a_j) conduisant au max $c(a_k,a_\ell)$; celui-ci correspond au max $\varphi(a_k,\bar{a}_\ell)$. En cas de solutions multiples, on

choisit le couple (a_i,a_j) tel que n_{a_i} soit le plus petit effectif parmi tous les effectifs des variables figurant en 1ère place des couples ; si à nouveau il y a solutions multiples, on choisit a_j tel que n_{a_j} soit l'effectif minimum parmi tous les effectifs des variables figurant en 2ème place des couples. En cas de nouvelle équivoque, on choisit l'un quelconque des couples restants.

Ainsi, $\forall(k,\ell),\varphi\,(a_i,\bar{a}_j) \geq \varphi\,(a_k,\bar{a}_\ell)$, et par suite : $\forall(k,\ell),c\,(a_i,a_j) \geq c\,(a_k,a_\ell)$, et on agrège l'ensemble $\{a_i,a_j\}$ en la classe (a_i,a_j) orientée de a_i vers a_j. Notons que, bien que $n_{a_i} \leq n_{a_j}$, n_{a_i} n'est peut-être pas l'effectif minimum pour l'ensemble des variables.

2ème étape : on compare les cohésions obtenues selon les phases $\mathcal{O}_2$ et $\mathcal{O}_3$ suivantes et on conserve la classe correspondant à la plus forte.

$\mathcal{O}_2$: on détermine le couple (ou l'un des couples) optimal dont la cohésion est immédiatement inférieure ou égale à la cohésion retenue à l'étape précédente ;

$\mathcal{O}_3$: on détermine l'ensemble $\{a_i,a_j,a_k\}$ puis la classe associée (au sens défini p. 18) dont la cohésion est meilleure que l'une quelconque des cohésions des couples restant après l'étape précédente.

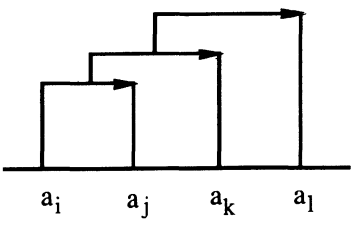
Notons alors que $c(a_i,a_j) \geq C((a_i,a_j),a_k)$.

Cette propriété tient à la définition de la cohésion comme moyenne géométrique et à la construction précédente de la classe (a_i,a_j) :

$$c\,(a_i,a_j) \geq [c(a_i,a_j)c\,(a_j,a_k)c\,(a_i,a_k)]^{\frac{1}{3}}$$

Finalement, lors de cette étape, on agrège une paire ou une classe à 3 éléments, mais simultanément, la cohésion de la nouvelle classe n'est pas meilleure que lors de l'étape précédente.

3ème étape : on compare les cohésions obtenues selon les phases $\mathcal{O}_2$, $\mathcal{O}_3$ et $\mathcal{O}_4$ suivantes et on conserve la classe correspondant à la plus forte.

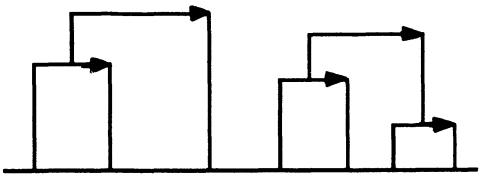


$\mathcal{O}_2$: on détermine un couple "*maximal*" parmi les couples restants ;

$\mathcal{O}_3$: on compare la cohésion de ce couple à celle des classes à 3 éléments non encore tous réunis, mais dont 2 d'entre eux l'ont déjà été ;

$\mathcal{O}_4$: on compare la cohésion maximale des 2 phases à celle des classes à 4 éléments élargissant la classe à 3 éléments éventuellement constituée lors de la 2ème étape en réunissant les 2 classes à 2 éléments déjà formés. Si tel est le cas avec $\{a_i,a_j,a_k,a_\ell\}$,
alors : $C((a_i,a_j),a_k) \geq C(((a_i,a_j),a_k),a_\ell)$
et a fortiori : $c(a_i,a_j)) \geq C(((a_i,a_j),a_k),a_\ell)$.

L'argument est le même que précédemment (cf. moyenne géométrique). La cohésion obtenue à l'issue des phase$\mathcal{O}_2$, $\mathcal{O}_3$ et$\mathcal{O}_4$ n'est pas meilleure que la précédente.
pième étape: avant cette étape, ont été agrégées éventuellement des classes à :



. 2 éléments, construites pas à pas selon une cohésion décroissante ;
. 3 éléments, construites pas à pas selon une cohésion décroissante et non meilleure que celle des couples les ayant générées ;
. 4 éléments, construites pas à pas selon une cohésion décroissante et non meilleure que celle des classes à 3 éléments les ayant générées ;
. .....
. p éléments selon les mêmes critères.

Selon les mêmes phases que précédemment, si une classe issue d'un ensemble à $(m+1)$ éléments est constituable à cette étape, c'est qu'elle réalise une cohésion maximale parmi toutes celles à $q \leq m$ éléments constituables à cette étape, compte tenu de celles déjà constituées. De plus, la cohésion de cette classe à $(m+1)$ éléments n'est pas meilleure que celle de la classe à m éléments qu'on veut étendre. (Si une classe à $m+r$ éléments est constituable, l'argument est valide a fortiori).

En conclusion, d'une étape à l'autre, la cohésion des classes formées est toujours décroissante au sens large. De plus, le processus est fini puisqu'à chaque étape, on accroît l'effectif d'une classe d'au moins un élément qui ne sera plus isolé. Ce processus prend fin lorsque toute nouvelle classe constituable admet une cohésion nulle ou bien lorsque la classe ultime est constituée de tous les éléments.

Un doctorant, sous la direction de R.GRAS, Saddo AG ALMOULOUD, élabore et étend sur P.C. et à partir de petites réalisations informatiques déjà opérationnelles, des programmes d'analyses de données appelés CHIC ou Classification Hiérarchique, Implicative et Cohésitive incluant :
. le calcul d'intensités d'implication entre attributs ou variables binaires et entre variables modales et variables numériques ;
. le calcul des cohésions et d'indices d'implication entre classes selon l'algorithme précédent ;
. la construction de l'arbre d'implication entre classes sur le modèle algorithmique élaboré par R. GRAS pour l'arbre de classification hiérarchique selon l'A.V.L. de I.C. LERMAN.

Ces programmes ont servi aux traitements de données dont il sera question ultérieurement. Nous renvoyons à la thèse pour de plus amples informations sur les logiciels en question.

2.4.2. Consistance de classe et minimalité.
On dira qu'à un niveau donné (lors d'une étape donnée), une classe $\mathcal{C}$ ***consistante*** vient de se former par agrégation de la classe $\mathcal{C}_2$ à $\mathcal{C}_1$ (éventuellement singleton) s'il existe au moins une classe $\mathcal{C}_0$ telle que la relation implicative entre classes soit améliorée :

$\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$

$\psi(\mathcal{C}_1, \mathcal{C}_0) < \psi(\mathcal{C}_1 \cup \mathcal{C}_2, \mathcal{C}_0)$ ou $\psi(\mathcal{C}_0, \mathcal{C}_1) < \psi(\mathcal{C}_0, \mathcal{C}_1 \cup \mathcal{C}_2)$.

De plus, soit $\mathcal{C}'$ une extension quelconque de $\mathcal{C}$.

Si $\forall\, \mathcal{C}_0 \not\subset \mathcal{C}$, $\psi(\mathcal{C}_0, \mathcal{C}') \leq \psi(\mathcal{C}_0, \mathcal{C})$

ou $\psi(\mathcal{C}', \mathcal{C}_0) \leq \psi(\mathcal{C}, \mathcal{C}_0)$,

$\mathcal{C}$ est dite *consistante minimale* (elle est minimale par rapport à son effectif d'éléments constituants).

Sous la même direction que les précédents doctorants, Harrisson RATSIMBA-RAJOHN étudie dans sa thèse la notion de noeud significatif lors de la construction de la hiérarchie, de

même que la contribution d'un individu ou d'une catégorie d'individus à la constitution d'une classe de variables.

En conclusion, l'implication statistique entre variables, prolongée en implication entre classes de variables nous fournit un outil d'étude de caractères de nature très variée. Elle semble utile au didacticien, au psychologue et de façon générale à tout chercheur disposant de données où les liaisons dans une population sont floues mais structurables en arbre puis en classes orientées. Des questions en termes de genèse, de complexité, de conduites nécessaires, etc., peuvent y trouver réponse. Nous cherchons à le montrer dans le paragraphe suivant.

3. APPLICATIONS DE L'IMPLICATION STATISTIQUE A L'ANALYSE DIDACTIQUE DE QUESTIONNAIRES.

L'étude présentée ici vise à connaître l'origine et la nature des erreurs les plus fréquentes, à analyser les procédures utilisées par des élèves de 1^{er} cycle mis en situation de démonstration en géométrie, en particulier dans le cas où l'activité déductive se réduit à une simple inférence. Nous entreprenons, sur des situations réelles - réponses d'élèves à un questionnaire -, les traitements statistiques des données recueillies, suivant les deux méthodes d'analyse présentées dans le § 1 : la classification hiérarchique (selon I.C. LERMAN) et la classification implicative. Nous dégageons ainsi quelques grandes classes de comportements erronés entre lesquelles nous tentons de nouveau d'établir des relations implicatives suivant le processus élaboré et développé dans le § 2.

3.1. Présentation du questionnaire.

Ce questionnaire concerne le début de l'apprentissage de la démonstration (classe de $5^{ème}$). Il est réalisé à partir d'un logiciel : le logiciel "PREMIER PAS", conçu au sein de l'équipe de didactique de Rennes pour aider l'enseignant à repérer et analyser les erreurs commises par l'élève dans les démonstrations à un pas. Ce logiciel propose à l'élève une liste de faits et une liste de théorèmes repérés par des numéros. Les questions concernent des inférences simples : hypothèse(s) - théorème - conclusion, présentant une ou plusieurs lacunes où sont fournis les numéros des faits ou théorèmes appropriés.

Le questionnaire, appelé "6 questions" que nous étudions ici, se rapporte à la symétrie centrale et, pour une de ces questions (Q_5), à la transitivité du parallélisme. Les questions posées visent à étudier l'effet d'un certain nombre de variables liées à la logique de l'inférence et à la forme des énoncés proposés. Hypothèses et théorèmes sont donnés : l'élève doit compléter en choisissant un des faits de la liste à titre de conclusion.

FAITS
1 (EF) et (CD) sont symétriques par rapport au point I
2 [MN] est le symétrique de [PR] par rapport au point I
3 (AB) et (CD) sont symétriques par rapport au point 0
4 (MN) // (PR)
5 (CD) // (EF)
6 (AB) // (CD)
7 (AB) // (EF)
8 MN = PF
9 CD = EF
10 AB = CD
11 AB = EF

THÉOREMES

1 La symétrie centrale conserve les longueurs.
2 Si (D) // (D') et (D') // (D"), alors (D) // (D").
3 Le symétrique d'une droite (D) par rapport à un point est une droite (D') parallèle à (D).
4 Si deux droites sont symétriques par rapport à un point, alors elles sont parallèles.
5 Deux segments symétriques par rapport à un point ont même longueur.
6 La symétrie centrale conserve les directions.

En fait, chaque question se présente schématiquement ainsi :
Hypothèse : fait n° p *Théorème* : n° q *Conclusion* : fait n° ?

Schématiquement, l'ensemble questions-réponses peut être présenté ainsi :

	HYPOTHESES	THEOREME	? CONCLUSION à trouver
Q1 Hypothèse : 1 Théorème : 3 Conclusion : 5	(EF) et (CD) symétriques par rapport à I	Le symétrique de (D) par rapport à un point est (D') // (D)	(EF) // (CD)
Q2 Hypothèse : 3 Théorème : 4 Conclusion : 6	(AB) et (CD) symétriques par rapport à 0	Si 2 droites sont symétriques par rapport à un point, alors elles sont parallèles	(AB) // (CD)
Q3 Hypothèse : 2 Théorème : 5 Conclusion : 8	[MN] est symétrique de [PR] par rapport à I	2 segments symétriques par rapport à un point ont même longueur	MN = PR
Q4 Hypothèse : 3 Théorème : 6 Conclusion : 6	(AB) et (CD) symétriques par rapport à 0	La symétrie centrale conserve les directions	(AB) // (CD)
Q5 Hypothèse : 6 et 5 Théorème : 2 Conclusion : 7	(AB) // (CD) et (CD) // (EF)	Si (D) // (D') et (D') // (D"), alors (D) // (D")	(AB) // (EF)
Q6 Hypothèse : 2 Théorème : 1 Conclusion : 8	[MN] est symétrique de [PR] par rapport à I	La symétrie centrale conserve les longueurs	MN = PR

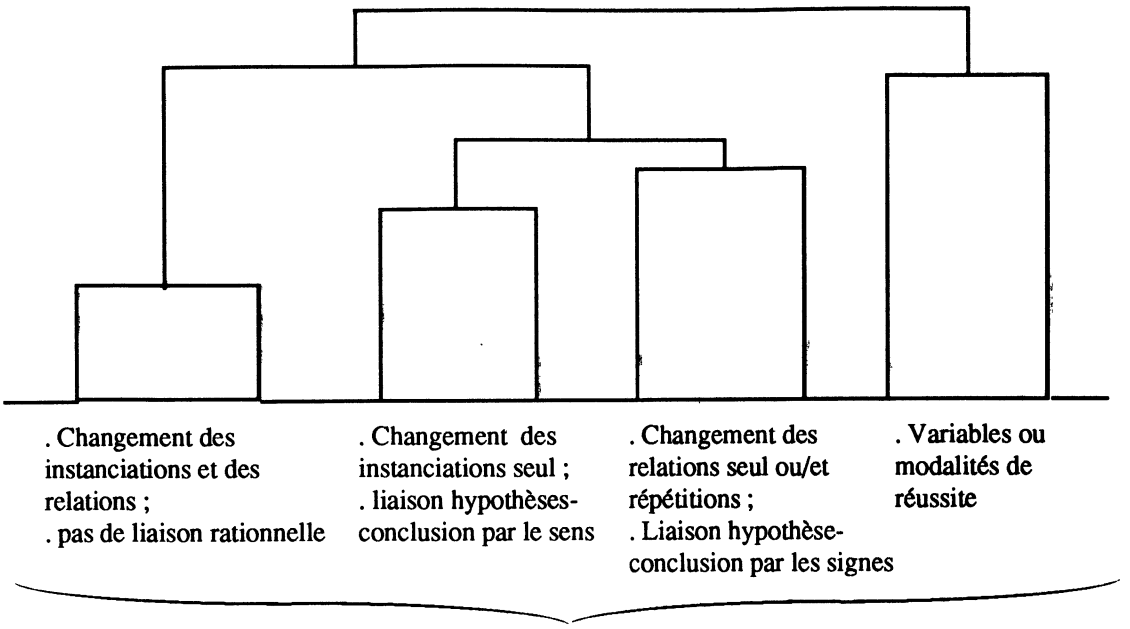
A la suite du 1er essai, l'élève est autorisé à faire un 2ème et dernier essai. Les questions sont indépendantes.

Chaque modalité de réponse est codée par un triplet.

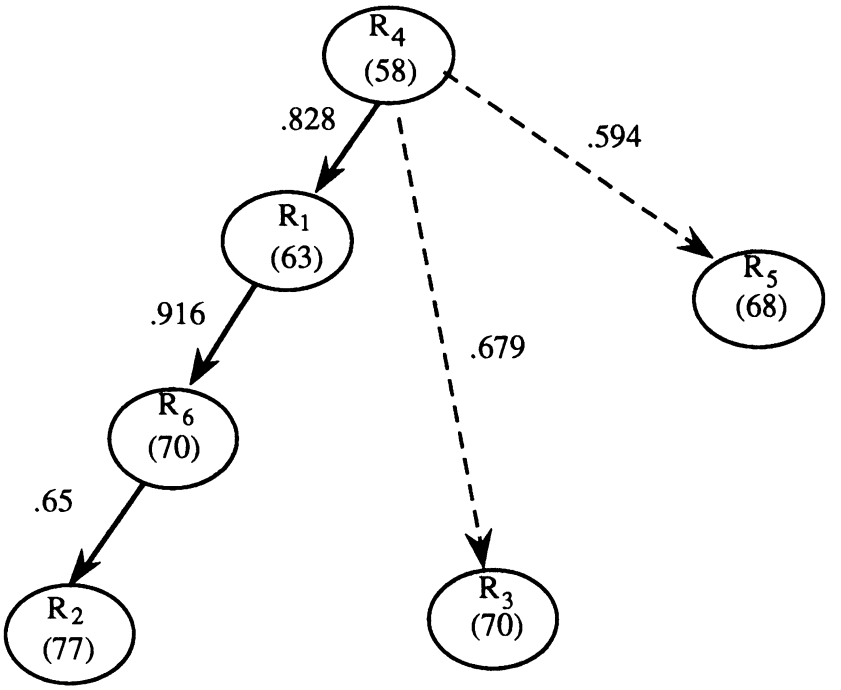
Exemple. 3-6-10 (Q_4).
Hypothèse : (*AB*) et (*CD*) sont symétriques par rapport à 0.
Théorème : la symétrie centrale conserve les directions.
Conclusion donnée par l'élève : *AB* = *CD*.

3.2. Analyse hiérarchique et implicative des réponses au "6 questions".

Après calcul des indices de similarité entre les 31 modalités de réponse prises 2 à 2 (80 élèves de 5ème), nous obtenons un arbre hiérarchique selon l'indice de similarité de I.C. LERMAN, que nous schématisons de la façon suivante :



Le *graphe implicatif* des réussites du "6 questions" traduit une relation de préordre partiel dans cet ensemble. Il est donc orienté et valué. R_4, réussite à la question (Q_4) la plus complexe du questionnaire (et la moins bien réussie), en est la *source*. R_2, R_3,R_5 en sont les puits :



Ci-contre, nous présentons le graphe implicatif entre les 6 modalités de réussite

R_2 , réussite à la question (Q_2) dont le théorème est exprimé en "si .... alors", formulation facilitant, semble-t-il, le succès.

R_3 , relative à la conservation des longueurs dans la symétrie centrale. La longueur est une notion plus familière à l'élève que celle de direction.

R_5 , seule question relative à la transitivité du parallélisme.

Les valeurs indiquées sont les intensités des implications.

Sur le même chemin, les réussites se placent dans l'ordre croissant de leurs effectifs.

Ainsi, dans le "6 questions", $R_1 \Rightarrow R_2$ avec une intensité d'implication très forte : 0,938. Les questions (Q_1) et (Q_2) ne diffèrent que par l'expression de leur théorème. Celui de (Q_2), en si....alors, facilite la réussite à cette question.

$R_4 \Rightarrow R_5$ très faiblement, avec une intensité d'implication de 0,594 ; R_5 est la seule question non relative à la symétrie centrale et contenant une double hypothèse.

3.3. Analyse des modalités de réponses au questionnaire "6 questions" (52 élèves de 5ème).

Question par question, nous avons identifié et codé les procédures susceptibles d'apparaître. Nous les présentons ci-dessous :

A : bonne réponse
R : répétition

Changement d'instanciation
- I_1 : changement total d'instanciation
- I_2 : changement partiel d'instanciation

Confusion entre objets
- O_1 : confusion entre droites et segments
- O_2 : confusion entre droites et longueurs
- O_3 : confusion entre segments et longueurs

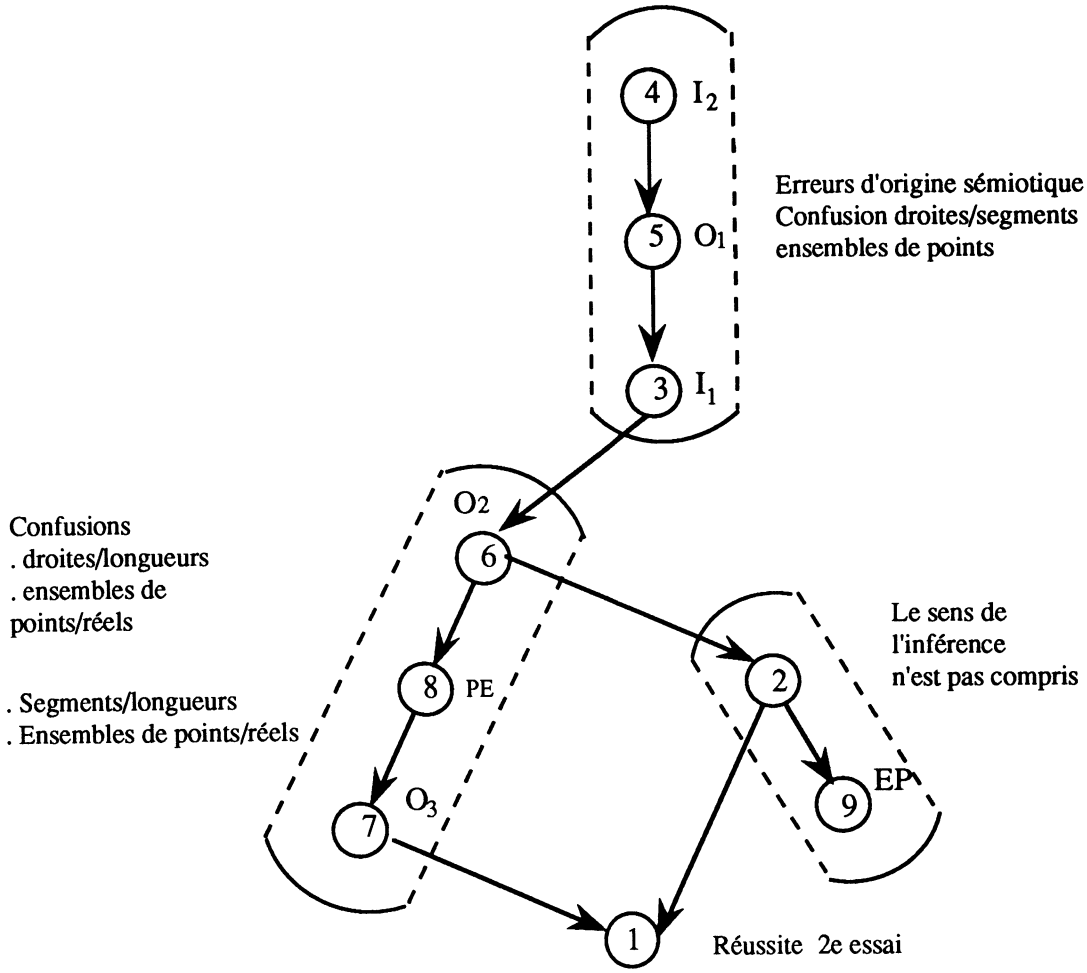
Confusion entre relations
- PE : // remplacé par =
- EP : = remplacé par // .

Ces procédures sont spécifiques du questionnaire et permettent donc de définir les compteurs non-standard de notre analyse.

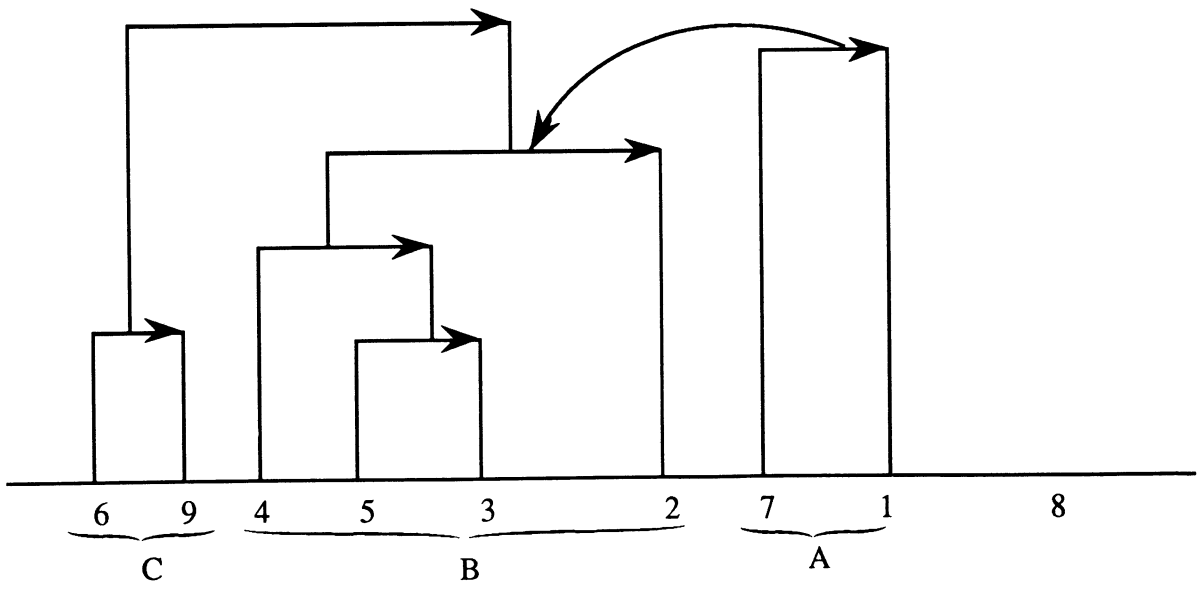
Variables	Nombre maximal d'occurences chez un élève	Occurences des variables sur les 52 élèves
1. Nombre de bonnes réponses au 2ème essai	5	66
2. Répétition	10	70
3. Changement total d'instanciation	10	29
4. Changement partiel d'instanciation	10	16
5. Confusion : droites/segments	6	10
6. Confusion : droites/longueurs	10	70
7. Confusion : segments/longueurs	4	49
8. // remplacé par = (PE)	6	64
9. = remplacé par // (EP)	4	31

Nous avons construit (cf. 1.3) une méthode permettant de donner du sens et d'attacher une intensité à la quasi-implication d'une variable numérique ou fréquentielle sur une autre.

Ainsi, le graphe implicatif obtenu après deux essais des élèves est le suivant :



Les calculs de cohésion et l'application de l'algorithme de constitution de classe nous permettent d'obtenir la classification suivante :



Elle partitionne l'ensemble des 9 variables en 3 classes disjointes, ce qui n'est pas le cas de la classification hiérarchique des similarités où l'emboîtement se produit dès le 5ème niveau.

La classe centrale B, constituée des deux sous-classes (4,(5,3)) et (3) met en évidence l'origine majeure des erreurs au 1er essai : le changement d'instanciation, la confusion droites/segments et la répétition, erreur qui peut comme les deux premières ne pas être persistante. Est encore moins persistante l'erreur de notation représentée par la variable 7 puisque celle-ci se referme avec 1 (réussite au 2ème esai) en une classe A très cohérente.

Remarquons le regroupement C des 2 variables 6 et 9 qui, par contre, renforce notre hypothèse d'erreur persistante car conceptuelle. La résistance à l'association de 8 à (6,9), révélée par la très faible cohésion, souligne, à son tour, la disjonction des procédures représentées par 8 (// remplacé par =) et 9 (= remplacé par //). En cela, la hiérarchie implicative nous informe de façon différente de la hiérarchie des similarités, relativement linéaire ici dans ses emboîtements et, par suite, moins pertinente pour l'analyse cognitive que nous avons menée.Ce jugement n'affecte en rien les qualités informationnelles de la méthode dans d'autres cas, comme nous l'avons constaté sur l'arbre des similarités du questionnaire.

3.4. Synthèse

Les analyses précédentes ont mis en évidence, dans le questionnaire, des types d'erreurs faites par de jeunes élèves en situation d'apprentissage de la démonstration mathématique ; 3 grandes familles de procédures erronées, très stables, mais non disjointes, se sont dégagées :
- changement d'instanciation : l'élève change les noms des objets
- répétition : l'élève répète en conclusion l'hypothèse de l'inférence
- changement de relation par rapport à l'attendu du théorème ou, plus spécifiquement, confusion entre le parallélisme et l'égalité.

L'examen des classifications hiérarchiques des similarités et des implications conduit à une épistémologie artificielle de l'inférence :
- *deux formes primitives de l'inférence* :
  . la répétition ou tautologie inféconde,
  . le changement des noms des objets avec changement ou non de la relation attendue ;
- *une forme plus évoluée* où les relations attendues sont échangées avec des relations voisines, sans symétrie entre ces échanges ;
- *une forme presque achevée* où la réponse diffère de l'attendu par la seule écriture (confusion entre signifiants seuls) ;
- *une forme achevée* conduisant à la réussite.

4. CONCLUSION.

Succédant à des mises à l'épreuve depuis plus de dix ans dans différentes recherches en didactique et en psychologie, les résultats obtenus à travers ces différents questionnaires, contrôlés localement par d'autres méthodes d'analyses de données (analyse factorielle et analyse hiérarchique), confortent l'approche implicative que nous avons adoptée. Par son caractère non symétrique, elle apporte une nouvelle synthèse dynamique de données, tout en complétant avantageusement les informations fournies par d'autres méthodes. L'extension à l'analyse implicative de classes présente un intérêt majeur par sa capacité à condenser les relations plus fines mais trop enserrées dans un réseau complexe. Des problèmes théoriques restent encore en suspens. Nul doute que l'examen des lois des intensités d'implication entre classes apportera une très intéressante information sur la richesse des liaisons. C'est, en particulier, l'objet de nouvelles autres recherches de notre équipe.

RÉFÉRENCES

[ACID S., de CAMPOS L.M., GONZALEZ A., MOLINA R., PEREZ de la BLANCA N. 1991] - Learning with Castle - in Symbolic and quantitative approaches to uncertainty (R. KRUSE, P. SIEGEL), Springer-Verlag, 99-106.

[AMARGER S., DUBOIS D., PRADE H. 1991], Imprecise quantifiers and conditional probabilities - in Symbolic and quantitative approaches to uncertainty (R. KRUSE, P. SIEGEL), Springer-Verlag, 33-37.

[DIDAY E. 1991] - Towards a statistical theory of intentions for knowledge analysis, rapport de recherche 1494, INRIA Rocquencourt.

[GAMMERMAN A., LUO Z. 1991] - Constructing Causal Trees from a medical database, Technical Report TR 91 002, Dep^{t} of Computer Sci., Heriot-Watt Univ., Edimburgh.

[GANASCIA J.G. 1991] - CHARADE : Apprentissages de bases de connaissances dans "Induction symbolique - numérique à partir de données", Ed. KODRATOFF et DIDAY, CEPADUES, 1991.

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

[GRAS R. et LARHER A., 1989] - La quasi-implication : une méthode d'analyse de relations non symétriques entre attributs et entre classes d'attributs, Public. interne I.R.M.A.R., Rennes, 1989.

[GUIGUES J.L. et DUQUESNE V. 1986] - Familles minimales d'implications informatives résultant d'un tableau de données binaires, Mathématiques et Sciences Humaines n° 95, p. 5-18, 1986.

[LARHER A., 1991] - Implication statistique et applications à l'analyse de démarches de preuve mathématique, Thèse de l'Université de Rennes I, février 1991.

[LERMAN I.C., GRAS R., ROSTAM H., 1981] - Elaboration et évaluation d'un indice d'implication pour des données binaires, I et II, Mathématiques et Sciences Humaines n° 74, p 5-35 et n° 75, p 5-47, 1981.

[LERMAN I.C., 1981] - Classification et analyse ordinale des données, Dunod, 1981.

[LOEVINGER J. 1947] - A systematic approach to the construction and evaluation of tests of ability, Psychological Monograhs, 61, n° 4.

[PEARL J. 1988] - Probabilistic Reasoning in intelligent systems, San Mateo, CA, Morgan Kaufmann.