MATHÉMATIQUES ET SCIENCES HUMAINES

ÉRIC TÉROUANNE

Une représentation graphique de la liaison statistique entre deux variables ordonnées

Mathématiques et sciences humaines, tome 130 (1995), p. 33-42

<http://www.numdam.org/item?id=MSH_1995__130__33_0>

Math. Inf. Sci. hum., (33e année, n°130, *1995*, pp.33-42)

UNE REPRÉSENTATION GRAPHIQUE DE LA LIAISON STATISTIQUE ENTRE DEUX VARIABLES ORDONNÉES

Éric TÉROUANNE[1]

RÉSUMÉ — *Le* stéréogramme de liaison *est une représentation graphique simultanée de la distribution conjointe de deux variables ordonnées, de leurs distributions marginales, et de la densité de la première par rapport au produit des deux autres. On y lit une forme de liaison statistique qui est introduite sous le nom de* liaison blackienne *et dont on discute les rapports avec la liaison stochastique.*

SUMMARY — A graphical representation of the statistical dependence between two ordered variables Dependence stereogram *is a simultaneous graphical representation of the joint distribution of two ordered variables, of their marginal distributions, and of the density of the former with respect to the product of the latter two. It is possible to read on it a form of statistical dependence which is introduced under the name of* blackian dependence *and whose connection with stochastic ordering is discussed.*

1. INTRODUCTION

L'étude de la liaison statistique entre deux variables ordonnées mesurées simultanément sur une population se ramène à celle d'un tableau de contingence dont les lignes et les colonnes sont totalement ordonnées *a priori*. Pour illustrer cet article, nous utiliserons les données de Tocher concernant la couleur des cheveux et la couleur des yeux de 5387 enfants (tableau 1), étudiées par Fisher (1940) et tiré de Rouanet & *al.* (1993, p.209). Dans chacun des ensembles de modalités, les couleurs sont ordonnées du plus clair au plus foncé.

Si l'on désigne par i une modalité de la première variable (ici une couleur d'yeux) et par j une modalité de la seconde (une couleur de cheveux), nous utiliserons les notations classiques : n_{ij} , $n_{i\cdot}$, $n_{\cdot j}$, $n_{\cdot\cdot}$ pour désigner les nombres d'individus (par exemple $n_{\text{clair,châtain}} = 584$; $n_{\cdot\text{noir}} = 118$; $n_{\cdot\cdot} = 5387$). De même, f_{ij} , $f_{i\cdot}$, $f_{\cdot j}$ désigneront les fréquences relatives correspondantes ($f_{xy} = n_{xy} / n_{\cdot\cdot}$) ; la fréquence conditionnelle $n_{ij} / n_{\cdot j}$ de la modalité i parmi les individus présentant la modalité j sera notée $f_{i/j}$.

[1] Département de Mathématiques et Informatique Appliquées, Université Montpellier III.

		Couleur des cheveux					
		Blond	Roux	Châtain	Marron	Noir	Total
Couleur des yeux	Pâles	326	38	241	110	3	718
	Clairs	688	116	584	188	4	1580
	Moyens	343	84	909	412	26	1774
	Foncés	98	48	403	681	85	1315
	Total	1455	286	2137	1391	118	5387

Tableau 1 - Distributions des effectifs conjoints et marginaux pour deux variables ordonnées.

La liaison éventuelle entre les deux variables se traduit en particulier dans les valeurs que prend la densité *d* de la distribution conjointe par rapport au produit de ses marginales. Celle-ci (tableau 2) est définie, pour tout couple (*i,j*) de modalités, par :

$$d_{ij} = \frac{f_{ij}}{f_{i\cdot}\, f_{\cdot j}} = \frac{n_{ij}\, n_{\cdot\cdot}}{n_{i\cdot}\, n_{\cdot j}}$$

		Couleur des cheveux				
		Blond	Roux	Châtain	Marron	Noir
Couleur des yeux	Pâles	1,68	1,00	0,85	0,59	0,19
	Clairs	1,61	1,38	0,93	0,46	0,12
	Moyens	0,72	0,89	1,29	0,90	0,67
	Foncés	0,28	0,69	0,77	2,01	2,95

Tableau 2 - Densité de la distribution conjointe par rapport au produit de ses marginales.

Divers indices sont reliés à cette densité, comme le taux de liaison entre une modalité de la première variable et une modalité de la seconde (*cf* par exemple Rouanet & *al.*, 1987, chapitre 7) défini pour tout couple (*i,j*) par :

$$t_{ij} = d_{ij} - 1$$

ou la contribution du couple (*i,j*) au ϕ^2 :

$$\phi_{ij}^2 = f_{i\cdot}\, f_{\cdot j}\, t_{ij}^2$$

Pour obtenir un indice de liaison global entre les variables, on peut alors calculer le ϕ^2, somme des ϕ_{ij}^2 , ou le $\chi^2 = n_{\cdot\cdot}\, \phi^2$. Mais ces indices, adaptés au cas de variables nominales, ne sont pas satisfaisants dans le cas qui nous intéresse puisqu'ils ne prennent pas en compte les structures d'ordre sur chacun des ensembles de modalités : le ϕ^2 et le χ^2 sont insensibles aux permutations de lignes ou de colonnes dans le tableau de contingence.

2 - STÉRÉOGRAMME DE LIAISON

Pour analyser la liaison entre deux variables en tenant compte des ordres *a priori*, il existe un outil graphique que nous appellerons le *stéréogramme de liaison.*

Le terme de stéréogramme est utilisé par Calot (*op.cit.* p.214) pour désigner un histogramme "tridimensionnel" représentant la distribution conjointe de deux variables numériques. Ces deux variables, les âges respectifs de l'époux et de l'épouse lors du mariage dans l'exemple de Calot, sont découpées en classes (ou modalités) d'amplitude variable. Chaque couple (i,j) de classes est alors représenté par un parallélépipède dont le volume est proportionnel à la fréquence conjointe f_{ij}. La longueur et la largeur de base du parallélépipède sont respectivement proportionnelles aux amplitudes a_i et a_j (en années dans l'exemple de Calot) des classes i et j. La hauteur du parallélépipède est ainsi proportionnelle à $f_{ij} / a_i a_j$.

Le stéréogramme de liaison (figure 1) représente également chaque paire de modalités (i,j) par un parallélépipède dont le volume est proportionnel à f_{ij} mais dont la base, contrairement au cas du stéréogramme classique, a une longueur et une largeur proportionnelles à $f_{i.}$ et $f_{.j}$ respectivement. La hauteur du parallélépipède est alors proportionnelle à la densité d_{ij}.

Cette représentation graphique possède donc cinq échelles différentes :

- Deux échelles linéaires pour les distributions marginales. La fréquence $f_{i.}$ d'une couleur d'yeux (resp. $f_{.j}$ d'une couleur de cheveux) est représentée par une portion du segment de longueur 1 figurant les abscisses (resp. les ordonnées).

- Une échelle de superficie pour représenter le produit des distributions marginales. Une fréquence $f_{i.}f_{.j}$ est représentée par la face supérieure du parallélépipède correspondant, c'est-à-dire une portion rectangulaire d'un carré de superficie totale égale à 1.

- Une échelle de volume pour représenter la distribution conjointe des deux variables. Une fréquence f_{ij} est représentée par le volume du parallélépipède correspondant. La somme de ces volumes est égale à 1.

- Une échelle de hauteur pour représenter la densité de la distribution conjointe par rapport au produit de ses marginales. La position sur cette échelle des faces supérieures des parallélépipèdes est représentée par des grisés (figure 1), et le taux de liaison t_{ij} se lit comme la différence entre la hauteur du parallélépipède correspondant et l'altitude 1 (en pointillé gras).

A chaque couleur de cheveux j (resp. à chaque couleur d'yeux i) est associée une "tranche" transversale du stéréogramme, constituée de quatre parallélépipèdes de même longueur qui correspondent aux quatre couleurs d'yeux (resp. cinq parallélépipèdes de même largeur correspondant aux cinq couleurs de cheveux).

3. LIAISON BLACKIENNE ENTRE VARIABLES ORDINALES

Rappelons la définition d'un ordre blackien (ou unimodal) par rapport à un autre (Black 1958, Barbut *et al.* 1971, Petit *et al.* 1988) : Soit ω un ordre total sur un ensemble X, pris comme ordre de référence. Soient ω' un autre ordre total sur X et x l'élément de X qui est maximal pour ω'. Alors ω' est dit blackien par rapport à ω s'il coïncide avec ω sur l'ensemble des éléments inférieurs ou égaux à x pour ω et avec le dual ω^* de ω sur l'ensemble des éléments

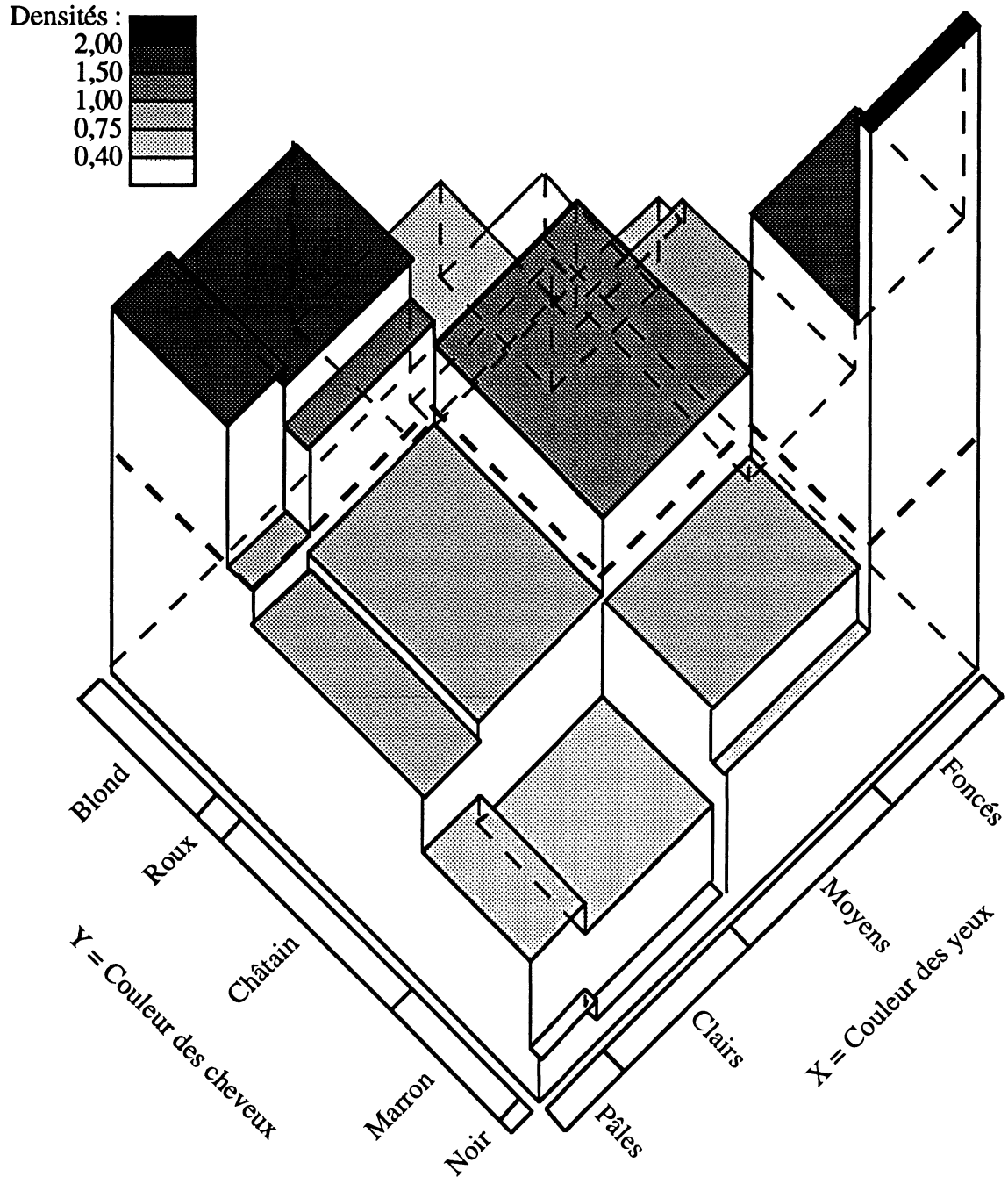

Figure 1. Une représentation graphique de la liaison statistique entre deux variable ordinales

supérieurs ou égaux à x pour ω. Autrement dit, ω' est obtenu en "pliant" l'ensemble X ordonné selon ω en un point x quelconque, et en réarrangeant les éléments des deux "brins" ainsi obtenus. S'il y a un seul brin, c'est-à-dire si x est le maximum (resp. le minimum) selon ω, le seul ordre ω' ainsi obtenu est ω (resp. ω*).

La liaison entre deux variables ordinales peut alors être caractérisée par une configuration particulière de leur distribution conjointe : Notons X l'ensemble des modalités de l'une (par exemple les couleurs d'yeux) et Y l'ensemble des modalités de l'autre (les couleurs de cheveux). Notons ω l'ordre *a priori* (dans notre exemple, du plus clair au plus foncé) sur X comme sur Y. Pour chaque modalité i de X, soient ω_i l'ordre obtenu sur Y en classant ses modalités selon la valeur des densités d_{ij}, et j_i la modalité maximale pour ω_i. Dans notre exemple (tableau 3) les j_i sont successivement Blond, Blond, Châtain, et Noir.

	Couleur des yeux			
	Pâles	Clairs	Moyens	Foncés
Maximum	Blond	Blond	Châtain	Noir
	Roux	Roux	Marron	Marron
Ordres des densités	Châtain	Châtain	Roux	Châtain
	Marron	Marron	Blond	Roux
Minimum	Noir	Noir	Noir	Blond

Tableau 3. Ordres des densités sur les couleurs de cheveux, à couleur d'yeux fixée.

DÉFINITION. On dira qu'il y a *liaison blackienne de la variable Y par rapport à la variable X* si tous les ordres ω_i sur Y sont blackiens par rapport à ω et que la suite des j_i est monotone. La liaison sera dite *croissante* (resp. *décroissante*) si la suite des j_i est croissante (resp. décroissante) vis-à-vis de ω. Enfin on dira qu'il y a *liaison blackienne réciproque entre les deux variables* s'il y a liaison de chacune par rapport à l'autre.

•

Il découle immédiatement du lemme ci-dessous que s'il y a liaison blackienne réciproque, les deux liaisons vont dans le même sens, croissant ou décroissant. Notons $x_0 \leq \ldots \leq x_i \leq \ldots \leq x_m$ les modalités de X et $y_0 \leq \ldots \leq y_j \leq \ldots \leq y_n$ celles de Y.

LEMME. S'il y a liaison blackienne croissante (resp. décroissante) de Y par rapport à X, alors :

$$d_{00} \geq d_{01} \text{ et } d_{mn} \geq d_{mn-1}$$
$$(\text{resp.} d_{0n} \geq d_{0n-1} \text{ et } d_{m0} \geq d_{m1}).$$

Supposons que la liaison soit croissante et que la première inégalité ne soit pas satisfaite, c'est-à-dire que $d_{00} < d_{01}$. On en déduit $j_0 > 0$ et, la suite des j_i étant croissante, on a :

$$j_i > 0 \text{ pour tout } i.$$

La suite des d_{ij} à i fixé étant blackienne, on en déduit :

$$d_{i0} < d_{i1} \text{ pour tout } i,$$

soit, en multipliant par $f_{i.}$:

$$f_{i0} < f_{i1} \text{ pour tout } i,$$

ce qui est impossible puisqu'on obtient 1 de part et d'autre de l'inégalité en sommant sur i.

Le même argument s'applique, *mutatis mutandis*, à la seconde inégalité ou au cas décroissant.

•

Comme le seul ordre blackien qui admette x_m (resp. x_0) comme plus grand élément est ω (resp. ω^*), on déduit immédiatement de ce lemme :

PROPOSITION. S'il y a liaison blackienne croissante de Y par rapport à X, l'ordre des densités associé à la plus grande modalité de X est l'ordre *a priori* sur Y et l'ordre des densités associé à la plus petite modalité de X est son dual.

•

Dans notre exemple la suite des j_i est croissante : il y a donc liaison blackienne croissante de la couleur des cheveux par rapport à la couleur des yeux. Par contre il n'y a pas une stricte liaison blackienne de la couleur des yeux par rapport à la couleur des cheveux : ω_{Marron} et ω_{Noir} ne sont pas blackiens par rapport à ω (tableau 4).

	Couleur des cheveux				
	Blond	Roux	Châtain	Marron	Noir
Maximum	Pâles	Clairs	Moyens	Foncés	Foncés
Ordres des densités	Clairs	Pâles	Clairs	Moyens	Moyens
	Moyens	Moyens	Pâles	Pâles	Pâles
Minimum	Foncés	Foncés	Foncés	Clairs	Clairs

Tableau 4. Ordres des densités sur les couleurs d'yeux, à couleur de cheveux fixée.

Cependant ces exceptions tiennent à de faibles différences entre les densités. Il suffirait de déclasser moins de 0,4% des individus, en en faisant passer 20 de (Pâles, Marrons) à (Clairs, Marrons) et un de (Pâles, Noirs) à (Clairs, Noirs), pour que ces interversions disparaissent. On pourrait donc dire qu'il y a "presque" liaison blackienne croissante réciproque entre les deux variables.

4.GRAPHISME

Graphiquement, la liaison blackienne réciproque entre deux variables se traduit sur le stéréogramme de liaison par le fait que les faces supérieures des parallélépipèdes (figure 1) forment un "col", autour d'un "point-selle", le couple (Châtain, Moyens) dans notre exemple.

Dans la mesure où la précision de l'échelle de grisés suffit, cette liaison se lit aussi bien sur une version simplifiée de la figure 1, un graphique bidimensionnel qui montre le stéréogramme de liaison "vu de dessus" (figure 2). On y retrouve la base carrée du stéréogramme, découpée en rectangles dont les longueurs et les largeurs sont proportionnelles aux fréquences marginales et les superficies proportionnelles aux fréquences produit. La présence d'un "col" s'y lit comme sur une carte hypsométrique.

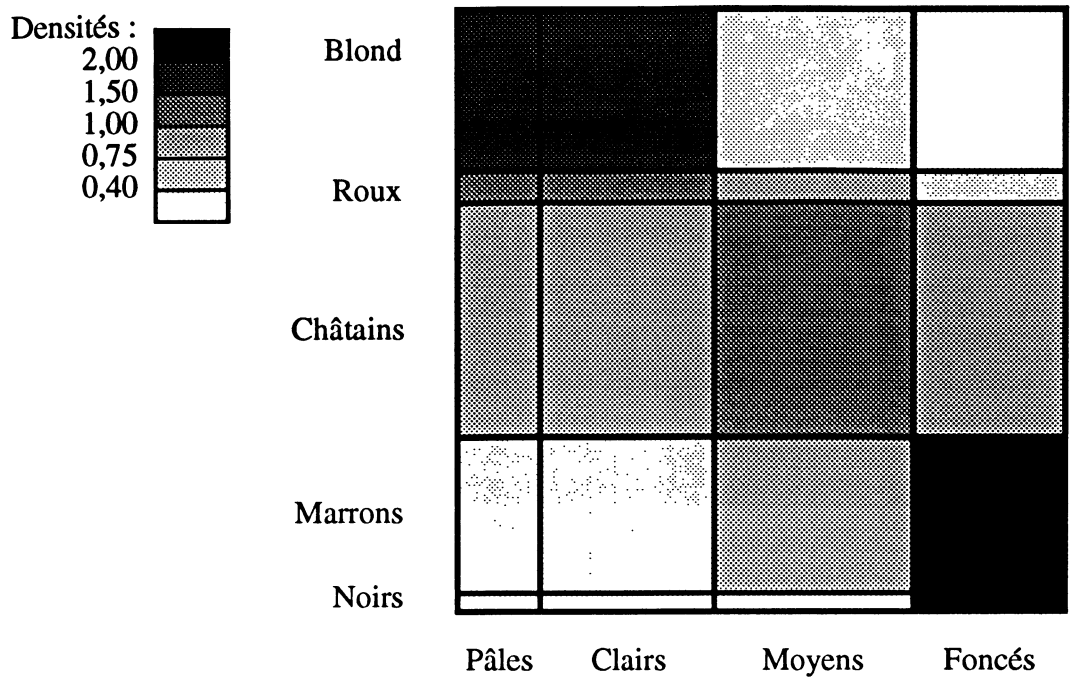

Figure 2. Vue à plat de la liaison blackienne[2].

Nous n'avons pas trouvé dans les outils graphiques proposés par les logiciels statistiques classiques le moyen de dessiner un stéréogramme (que ce soit celui de Calot ou le stéréogramme de liaison). Les graphiques "3D" qu'ils proposent utilisent en effet un quadrillage régulier en largeur comme en longueur. Il sera cependant possible d'utiliser ces outils pour dessiner un stéréogramme de liaison chaque fois que l'on pourra choisir des regroupements de modalités de façon à rendre les distributions marginales uniformes. Par exemple dans le cas de variables numériques découpées par leurs quartiles, déciles ou autres quantiles. En effet les fréquences f_{ij} de la distribution croisée seront alors directement proportionnelles aux densités d_{ij} et le stéréogramme de liaison se confondra avec la représentation "3D" de la distribution croisée.

5. LIAISON ENTRE VARIABLES NOMINALES

Le stéréogramme de liaison peut également servir à étudier la liaison entre des variables dépourvues d'ordre a priori. On peut par exemple chercher s'il existe un ordre sur X et un ordre sur Y qui donnent à la densité d une structure blackienne. Ou encore chercher des partitions de X et Y en deux ou plusieurs classes faisant apparaître des pics de densité au croisement d'une classe de X et d'une classe de Y, et des creux ailleurs. Cela peut se faire empiriquement en se guidant sur les valeurs inscrites dans le tableau des densités pour intervertir des lignes ou des colonnes, à la manière des "matrices ordonnables" de Bertin (*op. cit.*).

6. LIAISON BLACKIENNE ET DÉPENDANCE STOCHASTIQUE

L'ordre stochastique, ou ordre du cumul, permet de comparer deux distributions d'une variable ordonnée. On dit que la distribution f est "à gauche" de la distribution f' , et on note $f \leq f'$, si

[2] L'auteur remercie un rapporteur pour avoir suggéré l'ajout de la figure 2.

pour toute modalité i la fréquence cumulée F_i, somme des f_j pour $j \leq i$, est supérieure ou égale à F'_i.

Sur une représentation graphique (figure 3) où les modalités, représentées non par des points mais par des segments de longueur arbitraire (toute égales dans la figure 3), sont rangées en abscisse, de la plus petite à gauche à la plus grande à droite, la "courbe" qui représente F est alors entièrement au dessus et donc à gauche de celle qui représente F'.

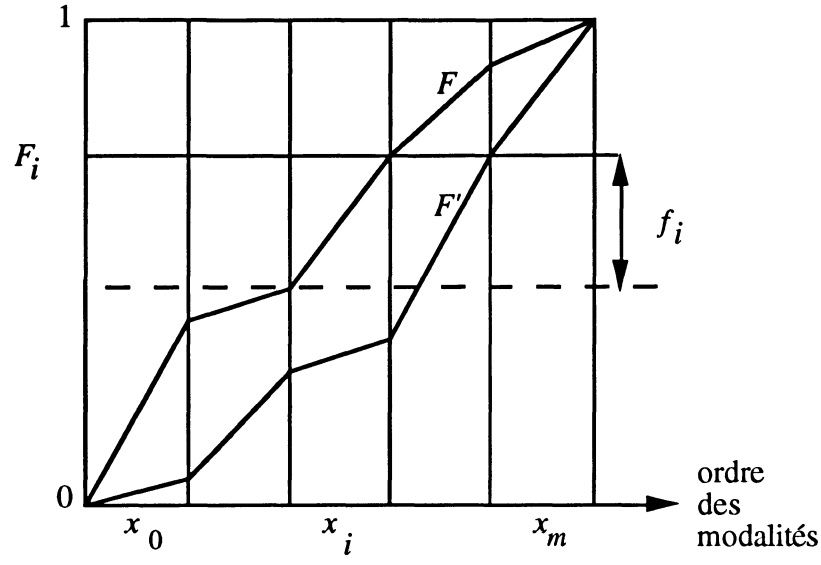

Figure 3. Ordre stochastique, ou ordre du cumul, entre deux distributions.

On associe à cet ordre sur les distributions une notion de dépendance statistique entre deux variables ordonnées X et Y : on dit que Y dépend stochastiquement de X de façon croissante si la famille des distributions de Y conditionnées par X est totalement ordonnée pour l'ordre du cumul :

$$i \leq i' \Rightarrow f_{/i} \leq f_{/i'}$$

Cette notion est décrite dans le cas de variables réelles sous le nom de "regression dependence" par Tukey (1958) et Lehmann (1959).

La relation entre dépendance stochastique et liaison blackienne est la suivante :

THÉORÈME. S'il y a liaison blackienne réciproque entre deux variables ordonnées X et Y, il y a également dépendance stochastique réciproque.

Notons $F_{j/i} = f_{0/i} + f_{1/i} + \ldots + f_{j/i}$ la fréquence cumulée de j conditionnée par i ($i \in X$, $j \in Y$). Considérons la représentation graphique de ces fonctions de répartition conditionnelles (figure 4) dans laquelle les modalités de Y sont représentées en abscisse par des segments de largeurs proportionnelles aux fréquences marginales f_j. Les pentes des segments successifs de la courbe des $F_{j/i}$ sont les densités d_{ij}. La liaison blackienne de Y par rapport à X se traduit donc sur ce graphique par le fait que chacune de ces courbes possède au plus un "segment d'inflexion", son segment de pente maximum (figuré en trait gras sur la figure 4), qui correspond au "pli" de l'ordre blackien, et par le fait que l'abscisse de ces segments se déplace de gauche à droite quand i croît de 0 à m.

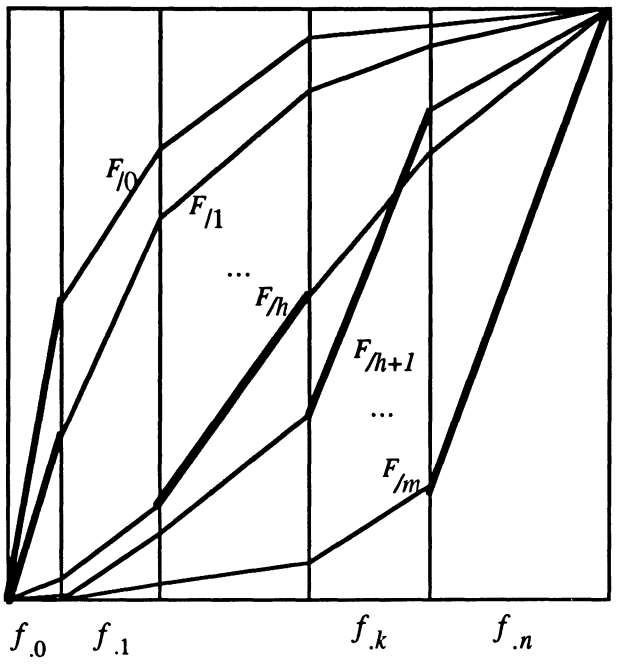

Figure 4. Fonctions de répartition conditionnelles de Y dans le cas d'une liaison blackienne croissante de Y par rapport à X.

On voit sur notre exemple (figure 4) que la liaison Blackienne de Y par rapport à X n'entraîne pas nécessairement la liaison stochastique : ici deux des courbes se croisent.

On suppose maintenant la liaison blackienne croissante réciproque. On doit démontrer :

$$F_{j/0} \geq F_{j/1} \geq \ldots \geq F_{j/n} \text{ pour tout } j.$$

Le lemme prouve que c'est vrai pour $j = 0$. Supposons le résultat vrai pour $j < k$ et faux pour k (pour un certain $0 < k < n$). Il existe une modalité h de x (figure 4) telle que :

$$F_{k-1/h} \geq F_{k-1/h+1} \quad \text{et}$$
$$F_{k/h} < F_{k/h+1}$$

Il s'ensuit que la pente de la courbe $F_{/h}$ au dessus de la modalité k est plus faible que celle de $F_{/h+1}$:

$$d_{hk} < d_{h+1,k}\,.$$

Comme les courbes finissent par se rejoindre, la courbe $F_{/h+1}$ présente au moins un segment, associé à une modalité $q > k$ de Y, dont la pente est strictement inférieure à la pente du segment correspondant de la courbe $F_{/h+1}$ ($q = n$ dans le cas de la figure 4). On a donc :

$$d_{hq} > d_{h+1,q}$$

Pour tout j fixé dans Y, la suite des d_{ij} est blackienne et atteint son maximum pour une modalité i_j dans X. Les inégalités ci-dessus entraînent :

$$i_k \geq h+1 \quad \text{et}$$
$$i_q \leq h\,, \quad \text{d'où}$$
$$i_q < i_k\,.$$

Ceci est en contradiction avec la croissance de la suite des i_j, puisque $k < q$. •

BIBLIOGRAPHIE

BARBUT M. et FREY L., *Techniques ordinales en analyse des données : Algèbre et combinatoire*, Paris, Hachette, 1971.

BERTIN J., *Sémiologie graphique,* Seconde édition, Paris, Mouton, 1973.

BLACK D., *The theory of committees and elections*, Cambridge, Mass, Cambridge Univ. Press, 1985.

CALOT G., *Cours de statistique descriptive*, Paris, Dunod, 1965, 2° éd. 1973.

FISHER R.A., "The precision of discriminant functions", *Ann. Eugen.*, 10, p.422-429, 1940.

LEHMANN E.L., *Testing statistical hypotheses*, New York, Wiley, 1959.

PETIT J.L. et TÉROUANNE E., *Résumons-nous,* Paris, Ellipses, 1988.

ROUANET H., LE ROUX B. et BERT M.C., *Statistique en Sciences Humaines : Procédures naturelles,* Paris, Dunod, 1987.

ROUANET H. et LE ROUX B., *Analyse des données multidimensionnelles*, Paris, Dunod, 1993.

TUKEY J.W., "A problem of Berkson, and minimum variance orderly estimators", *Ann. Math. Statist.,* 29, p.588-592, 1958.