MATHÉMATIQUES ET SCIENCES HUMAINES

ÉRIC TÉROUANNE

# Corrélation entre variables nominales, ordinales, métriques ou numériques

*Mathématiques et sciences humaines*, tome 142 (1998), p. 5-16

<http://www.numdam.org/item?id=MSH_1998__142__5_0>

*Math. Inf. Sci. hum.*, (36e année, n°142, *1998*, pp.5-16)

# CORRÉLATION ENTRE VARIABLES NOMINALES, ORDINALES, MÉTRIQUES OU NUMÉRIQUES

Éric TÉROUANNE[1]

RÉSUMÉ — *Un coefficient de corrélation est défini pour la distribution empirique conjointe de deux variables statistiques, que la structure a priori de chacune d'elles soit nominale, ordinale, métrique ou numérique. L'obtention d'un formalisme commun à toutes ces structures permet d'affiner l'analyse de la liaison entre les variables, en termes d'homogénéité (variables ordonnées), d'ordres sous-jacents (variables non-ordonnées) ou d'ordre induit (cas mixte).*

SUMMARY — Correlation between nominal, ordinal, metrical or numerical variables.
*A correlation coefficient is defined for the empirical joint distribution of two statistical variables, whatever be the a priori structure, nominal, ordinal, metrical or numerical of each other. Obtaining a common formalism for all these structures allows us to refine the analysis of correlation in terms of homogeneity (ordered variables), underlying orders (non-ordered variables) or induced order (mixed case).*

## 1. INTRODUCTION

Cet article est la troisième partie d'un travail entrepris afin de compléter et de présenter de façon cohérente un ensemble d'outils statistiques traitant de la comparaison de deux distributions en termes de tendance centrale ou en termes de dispersion, ou l'analyse de la liaison entre deux variables. Il se situe donc volontairement dans le champ de la statistique descriptive élémentaire. S'il lui arrive de rejoindre des travaux de portée beaucoup plus générale en analyse combinatoire des données (voir [1] pour une revue récente), il a pour ambition essentielle de combler les lacunes constatées par l'auteur dans la panoplie dont dispose habituellement le praticien en sciences humaines confronté à la simple description, non probabiliste, de situations statistiques uni- et bi-variées.

Avec l'outillage élémentaire le plus commun, on sait par exemple analyser la différence entre deux distributions d'une variable numérique, par la comparaison de leurs moyennes ou de leurs écarts-types, mais on ne dispose pas d'outils similaires dans le cas de variables ordinales ou nominales ; on sait mesurer la liaison entre deux variables numériques ou ordinales par divers coefficients de corrélation, mais on ne dispose pas d'instruments équivalents dans le cas

[1] Département de mathématiques et informatique appliquées, Université Paul Valéry, 34199 Montpellier Cedex 5. Couriel : EricTerouanne@bred.univ-montp3.fr

nominal ; enfin l'outillage commun laisse totalement de côté la large classe des variables dont la structure a priori est purement métrique. Ce dernier point nous amènera (§2) à proposer une typologie des «structures simples», réunissant les cas nominal, ordinal, métrique et numérique, qui sont pris en compte dans ce travail.

Dans une première étape ([9]), on a introduit la notion de «distorsion» comme mesure de la différence de tendance centrale entre deux distributions d'une variable nominale. On a ensuite montré ([10]) comment un formalisme commun, celui des «transferts», permet d'étendre cette mesure de la différence entre deux distributions aux quatre types de structures simples, en retrouvant les mesures classiques de la différence de tendance centrale dans le cas de deux structures particulières : le cas numérique, où le transfert coïncide avec la différence entre les moyennes des deux distributions, et le cas dichotomique, où il coïncide avec la différence absolue entre les fréquences de l'une ou l'autre des deux modalités.

Dans cette troisième étape, on suivra la même démarche pour parler de liaison entre variables et, au passage, de dispersion. On commencera par utiliser la notion de distorsion pour définir la «covariation» entre deux variables nominales (§3). On en tire un coefficient de corrélation nominal qui s'ajoute à une liste déjà longue de propositions concurrentes (cf [7] pour une revue récente), mais dont la justification sera d'apparaître comme cas particulier d'un indice commun à toutes les structures simples.

Pour parvenir à une mesure de la covariation commune à ces quatre types de structure (§4), on sera amené à formaliser la notion de «position» d'une modalité par rapport à une autre, puis celle de «co-position» d'un couple de modalités par rapport à un autre. On mettra ainsi en évidence une différence essentielle entre variables ordonnées (ordinales ou numériques) et variables non ordonnées (nominales ou métriques) : dans le premier cas la position ou la co-position ont un signe déterminé par la structure a priori, alors que dans le second on fait appel à un «signe par défaut», déterminé a posteriori par les distributions de fréquences que l'on veut comparer. On vérifiera que le coefficient de corrélation associé coïncide, dans chacune des structures particulières concernées, avec le coefficient de corrélation linéaire, le $\tau$ de Kendall ou le $\rho$ de Spearman, et on soulignera la singularité du cas numérique.

Pour finir on montrera comment l'avantage de posséder un formalisme commun aux variables ordonnées et non-ordonnées permet d'affiner l'étude de la liaison entre deux variables, avec la notion d'homogénéité dans le cas ordonné (§5), la notion duale d'ordres sous-jacents dans le cas non-ordonné (§6) ou encore celle d'ordre induit dans le cas mixte d'une variable ordonnée et une variable non-ordonnée (§7).

## 2. STRUCTURES SIMPLES

Devant toute variable statistique, le statisticien doit choisir la structure a priori qu'il désire prendre en compte sur l'ensemble $E$ de ses modalités. La nomenclature des structures possibles d'une variable statistique se réduit communément, dans les cours et ouvrages de statistique descriptive, à deux classes : qualitatives, quantitatives (cf. par exemple [2]), voire trois : nominales, ordinales, numériques (cf. par exemple [3]). De rares auteurs ([8]) mentionnent une quatrième structure : métrique, mais c'est pour la rattacher au cas numérique en se restreignant aux métriques euclidiennes de la droite, du plan ou plus généralement de $R^n$. Il faut aborder des travaux plus spécialisés, et encore hors de portée du grand public ([5], [1]), pour voir prise en compte la structure métrique d'une façon non triviale.

Or une part importante des variables rencontrées dans diverses disciplines d'observation, en particulier en sciences humaines ou en biologie, sont très naturellement munies par le spécialiste de structures métriques variées : ultramétrique associée à un arbre de classification (catégories socio-professionnelles), distance du plus court chemin sur un graphe (nœuds d'un réseau de communication), avec pour cas particulier les variables cycliques (temporelles ou spatiales), nombre de différences entre deux mots (génôme), etc. La structure nominale elle-même apparaît comme le cas particulier d'une structure métrique uniforme.

Aussi la structure métrique prend-elle naturellement sa place dans un tableau de quatre *structures simples* qui se proposent au choix du praticien (tab.1). Ce choix de la structure a priori se décompose alors en deux : celui de doter ou non $E$ d'un ordre total et celui de le doter ou non d'une distance (non uniforme). On dit que la variable est nominale si l'on ne choisit ni ordre ni distance, métrique si l'on ne choisit qu'une distance, ordinale si l'on ne choisit qu'un ordre.

| STRUCTURES SIMPLES | métrique uniforme | métrique non-uniforme |
|---|---|---|
| pas de relation d'ordre | (dont : DICHOTOMIQUE) NOMINALE | METRIQUE |
| relation d'ordre total | ORDINALE | METRIQUE & ORDRE (dont : NUMERIQUE) |

Tableau 1 - Schéma des quatre structures simples, avec deux cas particuliers.

La structure qui combine ordre et métrique admet comme cas particulier celui d'une variable numérique, pour laquelle on a choisi une distance et un ordre compatibles (tels en particulier que les inégalités triangulaires entre trois points ordonnés soient toutes des égalités). A l'opposé de ce cas particulier, que l'on peut qualifier de la plus «riche» des structures simples, on peut distinguer le cas le plus «pauvre», celui où le nombre des modalités est si faible que la question de la structure ne se pose même pas : le cas d'une variable dichotomique. Ce schéma laisse de côté nombre de structures plus complexes, comme les ordres partiels ou les structures produit, qui sortent du cadre de cette étude.

## 3. UN COEFFICIENT DE CORRÉLATION ENTRE VARIABLES NOMINALES

Pour mesurer la différence de tendance centrale entre deux distributions $p$ et $q$ d'une variable nominale, on a proposé ([9]) la notion de distorsion entre $p$ et $q$ définie par :

$$\delta(\{p,q\}) = \sum_{\{x,x'\}} |p(x)q(x') - p(x')q(x)|$$

Cette grandeur varie de 0, quand les deux distributions sont égales, à 1 quand elles sont étrangères (de supports disjoints).

Il paraît naturel d'élargir l'utilisation de cet indice à l'étude de la liaison entre deux variables. Soit en effet $P$ la distribution de fréquences conjointe de deux variables nominales $X$ et $Y$ observées simultanément sur une même population et à valeurs dans $E$ et $F$ respectivement. On notera $P(x,y)$ la fréquence conjointes de $(x,y)$, $P(x)$ et $P(y)$ les fréquences marginales de $x$ et de $y$, et $P_x$ (resp. $P_y$) désignera la distribution de $Y$ (resp. de $X$) conditionnelle à une modalité $x$ de $X$ (resp. $y$ de $Y$). Une influence éventuelle de $Y$ sur $X$ apparaîtra d'autant plus forte que les distributions conditionnelles $P_y$ seront plus différentes les unes des autres. On peut donc mesurer cette influence par la moyenne des distorsions $\delta(\{P_y, P_{y'}\})$ entre distributions conditionnelles de $X$, moyenne étendue à toutes les couples $(y,y')$ de modalités dans $F$:

$$\sigma(X,Y) = \sum_{y,y'} \delta(\{P_y, P_{y'}\})P(y)P(y')$$

Cette grandeur est nulle si et seulement si toutes les distorsions $\delta(\{P_y, P_{y'}\})$ sont nulles, c'est à dire si les distributions conditionnelles de $X$ sont toutes égales, ce qui revient à dire que $X$ et $Y$ sont indépendantes. A l'opposé, à distribution marginale de $Y$ fixée, $\sigma(X,Y)$ est maximale si toutes les distorsions $\delta(\{P_y, P_{y'}\})$ sont égales à 1, c'est à dire si les supports des distributions conditionnelles $P_y$ sont tous disjoints, ce qui signifie que $Y$ dépend fonctionnellement de $X$.

On peut de même évaluer l'influence éventuelle de $X$ sur $Y$ en calculant la distorsion moyenne entre les distributions conditionnelles $P_x$ de $Y$. On vérifie facilement que ces deux calculs donnent le même résultat, qu'on appellera *covariation* de $X$ et $Y$ et qui s'écrit, d'une façon plus symétrique :

$$\sigma(X,Y) = \sum_{\{(x,y),(x',y')\}} |P(x,y)P(x',y') - P(x,y')P(x',y)|$$

Appliqué à la distribution conjointe de la variable $X$ avec elle-même, cet indice mesure la *dispersion* de la distribution marginale de $X$. En effet chaque distribution conditionnelle $P_x$ se confond alors avec la distribution de Dirac en $x$, aussi la distorsion $\delta(\{P_x, P_{x'}\})$ vaut-elle 0 si $x = x'$ et 1 sinon, et l'on obtient :

$$\sigma(X,X) = \sum_{x \neq x'} P(x)P(x') = \sum_{x} P(x)(1-P(x)) = 1 - \sum_{x} P(x)^2$$

Cet indice est nul si et seulement si la distribution de $X$ est concentrée sur une seule modalité de $E$. Il est maximal (et égal à 1-1/Card($E$)) si et seulement si cette distribution est uniforme.

On mesurera naturellement la *corrélation* entre $X$ et $Y$ par :

$$\rho(X,Y) = \frac{\sigma(X,Y)}{\sqrt{\sigma(X,X)\sigma(Y,Y)}}$$

On vérifie aisément que ce coefficient de corrélation $\rho$ est toujours compris entre 0 et 1, qu'il est nul si et seulement si $X$ et $Y$ sont indépendantes et qu'il est égal à 1 si et seulement si les deux variables sont en dépendance fonctionnelle bijective. Dans le cas de deux variables dichotomiques, il coïncide (au signe près) avec le coefficient point-tétrachorique (voir par exemple [8]).

## 4. CORRÉLATION ENTRE DEUX VARIABLES DE STRUCTURE QUELCONQUE

### 4.1. Position

Le calcul du transfert entre une distribution $p$ et une distribution $q$ sur un même ensemble $E$ repose sur la notion sous-jacente de *position d'une modalité par rapport à une autre*. Cette position $\pi(x,x')$ est une fonction antisymétrique sur $ExE$ ($\pi(x,x')=-\pi(x',x)$), dont la valeur absolue $d(x,x')$ dépend de la distance choisie, et dont le signe $s(x,x')$ dépend de l'ordre choisi, de la façon suivante :

- La distance $d$ est la distance a priori sur $E$ si l'on en a choisi une, la distance uniformément égale à 1 en dehors de la diagonale sinon.

- Le signe $s(x,x')$ dépend de l'ordre a priori sur $E$ si l'on en a choisi un et des distributions $p$ et $q$ sinon. Si l'on a choisi l'ordre a priori $\omega$ sur $E$, $s(x,x')$ vaut +1 si $(x,x')$ appartient à $\omega$ (au sens de l'ordre strict) et -1 si c'est $(x',x)$ qui appartient à $\omega$ . Si l'on n'a pas choisi d'ordre a priori sur $E$, on lui substitue l'ordre de distorsion $(q/p)$ (cf[10]) : $s(x,x')$ vaut +1 si le déterminant $p(x)q(x')-p(x')q(x)$ est positif et -1 s'il est négatif. Dans tous les cas, $s(x,x)=0$.

On appelle alors *position de x par rapport à x'* la quantité $\pi(x,x')=s(x,x')d(x,x')$. Dans le cas d'une variable numérique, on retrouve bien sûr : $\pi(x,x') = x'-x$. Avec ces notations, le transfert $t(p,q)$ pour passer de la distribution $p$ à la distribution $q$ , tel qu'on l'a défini en [10], est, dans tous les cas de structure simple, la position moyenne pour le produit des distributions $p$ et $q$ :

$$t(p,q) = \sum_{(x,x')} \pi(x,x')p(x)q(x')$$

On note que la position de $x$ par rapport à $x'$ ne dépend que de la structure a priori quand celle-ci comporte un ordre, mais dépend également des deux distributions quand elle n'en comporte pas. C'est en cela qu'elle se distingue, dans le cas symétrique (sans ordre), de la notion utilisée par Lerman ([5]). Nous allons retrouver cette différence dans l'étude de la liaison entre deux variables.

4.2. Co-position

Soient donc $X$ et $Y$ deux variables définies sur une même population et à valeurs respectivement dans les ensembles $E$ et $F$ munis chacun d'une structure simple quelconque, et soit $P$ leur distribution conjointe. La *co-position* $\pi((x,y),(x',y'))$ d'un couple de modalités $(x,y)$ de $ExF$ par rapport à un autre couple $(x',y')$ est le produit d'une valeur absolue $d((x,y),(x',y'))$ et d'un signe $s((x,y),(x',y'))$ :

- La valeur absolue $d((x,y),(x',y'))$ est le produit $d(x,x')d(y,y')$ des distances entre $x$ et $x'$ d'une part, $y$ et $y'$ d'autre part, avec la convention que si l'on n'a pas choisi de distance sur l'un ou l'autre ensemble celle-ci est réputée uniforme.

- Le signe $s((x,y),(x',y'))$ vaut +1 ou -1. Il dépend des ordres choisis sur $E$ et $F$ s'ils existent tous deux, et de la distribution $P$ sinon. Dans le premier cas, $s((x,y),(x',y'))$ est le produit $s(x,x')\ s(y,y')$, qui vaut +1 si les couples $(x,y)$ et $(x',y')$ sont comparables pour l'ordre produit («$x<x'$ et $y<y'$» ou «$x'<x$ et $y'<y$»), -1 sinon («$x<x'$ et $y>y'$» ou «$x>x'$ et $y<y'$»). Dans le second cas, $s((x,y),(x',y'))$ vaut +1 ou -1 selon que le déterminant $P(x,y)P(x',y') - P(x,y')P(x',y)$ est positif ou négatif.

Le signe de $\pi$ dépend donc de la distribution $P$ si $E$ et $F$ ne sont pas tous deux ordonnés. L'antisymétrie de la co-position s'exprime par :

$$\pi((x,y),(x',y')) = \pi((x',y'),(x,y)) = -\pi((x,y'),(x',y)) = -\pi((x',y),(x,y')).$$

4.3. Covariation, dispersion et corrélation

La *covariation* entre $X$ et $Y$ est la co-position moyenne de tous les couples de modalités, pour le carré de la distribution $P$ :

$$\sigma(X,Y) = \sum_{(x,y)} \sum_{(x',y')} \pi((x,y),(x',y'))P(x,y)P(x',y')$$

Dans le calcul de la covariation entre une variable $X$ et elle-même, les sommes ci-dessus s'étendent à $ExE$, sur lequel la distribution $P$ a pour support la diagonale : $P(x,y) = P(x)$ si $x=y$, 0 sinon. Sur cette diagonale, $s((x,x),(x',x'))$ vaut toujours 1 si $x \neq x'$ : c'est en effet le carré de $s(x,x')$ si l'on a choisi un ordre a priori et le signe du déterminant $P(x)P(x')$ sinon. La covariation entre $X$ et elle-même s'écrit donc dans tous les cas :

$$\sigma(X,X) = \sum_{(x,x')} d(x,x')^2 P(x)P(x')$$

et sa racine carrée est la moyenne quadratique des distances entre modalités. On a vu au §3 la forme qu'elle prend dans le cas nominal. Elle prend la même dans le cas ordinal. C'est dans tous les cas un *indice de dispersion* de la distribution marginale de $X$, et le *coefficient de corrélation* entre $X$ et $Y$ est classiquement défini par :

$$\rho(X,Y) = \frac{\sigma(X,Y)}{\sqrt{\sigma(X,X)\sigma(Y,Y)}}$$

Ce coefficient s'applique donc au cas de deux variables de structures différentes, au prix de l'abandon de la structure d'ordre de l'une des variables si l'autre n'en a pas également reçu une. Il coïncide avec le coefficient défini au §3 dans le cas de deux variables nominales.

4.4. Corrélation entre deux variables non-ordonnées

La covariation entre deux variables métriques $X$ et$Y$ peut s'écrire :

$$\sigma(X,Y) = \sum_{\{(x,y),(x',y')\}} d(x,x')d(y,y') \left|P(x,y)P(x',y') - P(x,y')P(x',y)\right|$$

ou encore comme la moyenne du produit de la distance entre deux modalités de l'une des variables par le transfert entre les distributions conditionnelles corrrespondantes de l'autre :

$$\sigma(X,Y) = \sum_{x,x'} d(x,x')\ t(P_x,P_{x'})P(x)P(x')$$

$$= \sum_{y,y'} d(y,y')\ t(P_y,P_{y'})P(y)P(y')$$

4.5. Corrélation entre deux variables ordonnées

Dans le cas de deux variables ordonnées (et en particulier numériques), la co-position $\pi((x,y),(x',y'))$ est le produit des «positions marginales» $\pi(x,x')$ et $\pi(y,y')$ : $\sigma(X,Y)$ est donc le produit scalaire de ces deux fonctions réelles, définies sur $(ExF)^2$, dans la structure euclidienne associée à la distribution carrée $P^2$.

Dans le cas de deux variables numériques on retrouve la covariance $\sigma_{XY}$ habituelle, à un doublement près :

$$\sigma(X,Y) = \sum_{(x,y)} \sum_{(x',y')} (x' - x)(y'-y)P(x,y)P(x',y')$$

$$= \sum_{(x',y')} x'y'P(x',y') + \sum_{(x,y)} xyP(x,y) - \sum_{(x',y)} x'yP(x')P(y) - \sum_{(x,y')} xy'P(x)P(y')$$

$$= 2\left[\sum_{(x,y)} xyP(x,y) - \sum_{x} xP(x) \sum_{y} yP(y)\right] = 2\sigma_{XY}$$

Il s'ensuit que :

$$\sigma(X,X) = 2\sum_{x} x^2P(x) - 2\left[\sum_{x} xP(x)\right]^2 = 2\sigma_X^2$$

et donc que $\rho(X,Y)$ est le coefficient de corrélation linéaire entre X et Y.

Dans le cas de deux variables ordinales (avec la métrique uniforme), on a :

$$\sigma(X,Y) = 2 \sum_{x<x'} \sum_{y<y'} P(x,y)P(x',y') - P(x,y')P(x',y)$$

$$= \frac{2}{N^2}(n(\text{accords}) - n(\text{désaccords}))$$

où $N$ désigne la taille de la population (ou de l'échantillon) étudiée, $n$(accords) le nombre de paires d'individus pour lesquels l'ordre selon $X$ et l'ordre selon $Y$ coïncident et $n$(désaccords) le nombre de paires pour lesquels ils sont opposés. Ainsi la valeur normalisée correspondante, $\rho(X,Y)$, est bien l'une des généralisations possibles du coefficient de corrélation de Kendall.

Dans le cas de deux variables ordonnées munies de la métrique des rangs, on retrouve bien sûr le coefficient de corrélation de rangs de Spearman.

La covariation peut s'exprimer, dans tous les cas de variables ordonnées, comme la moyenne, pour la distribution marginale de l'une des variables, du produit des positions relatives de deux de ses modalités par le transfert entre les distributions conditionnelles correspondantes de l'autre variable :

$$\sigma(X,Y) = \sum_{x,x'} \pi(x,x')t(P_x,P_{x'})P(x)P(x')$$

$$= \sum_{y,y'} \pi(y,y')\, t(P_y,P_{y'})P(y)P(y')$$

Dans le cas de deux variables numériques, et si l'on note $Y_x$ la moyenne de la variable $Y$ pour la distribution conditionnelle $P_x$, on retrouve :

$$\sigma(X,Y) = 2\sigma_{XY} = \sum_{x,x'} (x'-x)(Y_{x'} - Y_x)P(x)P(x')$$

$$= \sum_{y,y'} (y'-y)(X_{y'} - X_y)P(y)P(y')$$

4.6. Singularité du cas numérique

Dans la présentation habituelle du coefficient de corrélation linéaire comme un cosinus, la covariance est définie comme un produit scalaire entre deux fonctions définies sur l'ensemble $\Omega$ des individus. La structure euclidienne utilisée sur l'espace de ces fonctions est le produit scalaire associé à une mesure a priori (par exemple uniforme) choisie sur la population $\Omega$. Cette définition admet comme cas particulier le $\rho$ de Spearman, cosinus entre les fonctions de rangs centrés. Par contre elle ne s'étend pas au $\tau$ de Kendall. Ce dernier possède bien une définition euclidienne, mais dans l'espace des fonctions définies sur les couples d'individus, et à valeur dans {-1,0,1} (voir par exemple [6], p.103-106).

C'est bien dans ce dernier cadre, l'espace des fonctions définies sur les couples d'individus, que s'écrit la définition de la covariation dans sa formulation générale. Dans le cas

particulier de deux variables numériques, ceci nous fournit une interprétation alternative du coefficient de corrélation linéaire : plutôt que le cosinus, dans $R^{\Omega}$, entre le vecteur des valeurs selon une variable et celui des valeurs selon l'autre, on le considère ici comme le cosinus, dans $R^{\Omega x\Omega}$, entre le vecteur des différences de valeur selon une variable et celui des différences de valeur selon l'autre.

Cette formulation commune au cas ordinal et au cas numérique fait apparaître ce dernier comme l'exception plutôt que la norme. La raison de cette exception est dans la simplification qui s'opère dans le cas numérique, du fait que $E$ possède lui-même une structure vectorielle : il existe alors une «moyenne» des modalités, et la position moyenne d'une modalité par rapport à toutes les autres se confond avec sa propre position par rapport à cette moyenne. Dans les autres structures, faute d'une telle simplification, les calculs demeurent en termes de positions relatives entre les modalités.

Cette nécessité, pour obtenir un langage commun à toutes les structures simples, de raisonner sur le carré cartésien de l'ensemble des modalités muni d'une distribution produit, et cette singularité du cas numérique, apparaissent dès la comparaison en termes de tendance centrale. En effet le transfert entre deux distributions $p$ et $q$ se définit comme une moyenne étendue à l'ensemble des couples d'individus (ou de modalités). Il admet la différence entre les moyennes d'une variable numérique comme cas particulier, et ce n'y a que ce dernier cas dans lequel son calcul se simplifie en se ramenant à l'ensemble des modalités ([10], p.70).

## 5. HOMOGÉNÉITE DE LA LIAISON ENTRE DEUX VARIABLES ORDONNÉES

On a introduit ([10]) la notion d'*homogénéité* du transfert entre deux distributions $p$ et $q$ d'une variable ordonnée : c'est le cas dans lequel l'ordre de distorsion $(q/p)$ coïncide avec l'ordre a priori $\omega$ (transfert homogène croissant) ou avec son dual $\omega^*$ (transfert homogène décroissant). Evaluer l'homogénéité du transfert revient à comparer le transfert mesuré dans la structure a priori, avec sa distance $d$ et son ordre $\omega$, au transfert mesuré dans la structure privée de cet ordre. Ces deux grandeurs peuvent s'écrire :

$$\sum_{(x,x')\in\omega} d(x,x') \begin{vmatrix} p(x) & q(x) \\ p(x') & q(x') \end{vmatrix} \quad \text{dans le cas ordonné}$$

$$\sum_{\{x,x'\}} d(x,x') \left\| \begin{matrix} p(x) & q(x) \\ p(x') & q(x') \end{matrix} \right\| \quad \text{dans le cas non-ordonné}$$

Le transfert ordonné est toujours, en valeur absolue, inférieur au transfert non-ordonné, et ne lui est égal que si le transfert est homogène. Le rapport entre la valeur absolue du premier et la valeur du second fournit ainsi un indice d'homogénéité, compris entre 0 et 1. On peut dire que si la valeur absolue du transfert mesure la force de la différence, l'indice d'homogénéité en mesure la qualité.

On vérifie donc l'homogénéité du transfert en s'assurant que pour tout couple $(x,x')$ appartenant à l'ordre $\omega$ le déterminant $p(x)q(x') - p(x')q(x)$ est positif. On montre facilement qu'il suffit en fait de le vérifier pour les couples où $x$ et $x'$ sont voisins dans l'ordre $\omega$. La démonstration se fait de proche en proche, à partir de l'étude de deux inégalités voisines, comme $p(x)q(x') > p(x')q(x)$ et $p(x')q(x'') > p(x'')q(x')$.

La même notion s'étend à l'étude de la liaison entre deux variables ordonnées : on dira que la liaison est *homogène croissante* (respectivement *décroissante*) si pour tout couple $((x,y),(x',y'))$ le signe $s((x,y),(x',y'))$ coïncide avec celui du déterminant $P(x,y)P(x',y') - P(x,y')P(x',y)$ (resp. avec le signe opposé). Evaluer l'homogénéité de la liaison revient donc à comparer la covariation mesuré dans la structure a priori, avec ses ordres $\omega$ sur $E$ et $\omega'$ sur $F$, au transfert mesuré avec les structures privées de leurs ordres. Ces deux grandeurs peuvent s'écrire :

$$\sum_{(x,x')\in\omega}\sum_{(y,y')\in\omega'} d(x,x')d(y,y')\begin{vmatrix} P(x,y) & P(x,y') \\ P(x,y') & P(x'y') \end{vmatrix} \quad \text{dans le cas ordonné}$$

$$\sum_{\{x,x'\}}\sum_{\{y,y'\}} d(x,x')d(y,y')\left\|\begin{matrix} P(x,y) & P(x,y') \\ P(x,y') & P(x'y') \end{matrix}\right\| \quad \text{dans le cas non-ordonné}$$

La covariation ordonnée est toujours, en valeur absolue, inférieure à la covariation non-ordonnée. Le rapport de la valeur absolue de la première à la seconde fournit donc un indice d'homogénéité de la liaison, compris entre 0 et 1. La liaison est homogène si et seulement si ce coefficient vaut 1. Ce type de dépendance a été introduit par Lehman ([4]) dans le cadre inférentiel de l'étude de variables aléatoires réelles, sous le nom de «likelihood ratio dependence». Cet auteur le présente comme le plus fort d'une série de concepts de dépendance étudiés dans ce même cadre.

On vérifie donc l'homogénéité de la liaison en s'assurant que pour toute paire de couples $(x,y)$ et$(x',y')$ comparables pour l'ordre produit, le déterminant $P(x,y)P(x',y') - P(x,y')P(x',y)$ est positif. Comme dans le cas de l'homogénéité du transfert, et en appliquant le même type d'argument au produit des ordres $\omega$ et $\omega'$, on montre qu'il n'est pas nécessaire pour établir l'homogénéité de la liaison entre deux variables ordinales de vérifier la positivité de tous les déterminants : il suffit de la vérifier pour les déterminants concernant des couples dont les modalités $x$ et $x'$, comme les modalités $y$ et $y'$, sont voisines dans l'ordre a priori.

Une autre condition équivalente à l'homogénéité croissante (resp. décroissante) est que pour tout couple $x<x'$ de modalités de l'une des variables, le transfert entre les distributions conditionnelles $P_x$ et $P_{x'}$ de l'autre variable soit homogène croissant (resp. décroissant).

L'intérêt de la notion d'homogénéité, comme des autres concepts introduits par Lehmann, est de proposer un diagnostic fin de la qualité de la liaison, indépendamment de sa force qui, elle, est mesurée par la valeur absolue du coefficient de corrélation.

## 6. ORDRES SOUS-JACENTS A UNE LIAISON ENTRE VARIABLES NON ORDONNÉES

De façon symétrique, l'homogénéité peut se «reconnaître» dans la distribution conjointe de deux variables non ordonnées (métriques ou nominales). C'est ce qui se passe s'il existe deux ordres $\omega$ sur $E$ et $\omega'$ sur $F$ tels que le signe du déterminant $P(x,y)P(x',y') - P(x,y')P(x',y)$ coïncide avec le produit des signes $s(x,x')$ associé à $\omega$ et $s(y,y')$ associé à $\omega'$. L'analyse de $P$ pour les structures de $X$ et $Y$ enrichies par ces ordres montre en ce cas une liaison homogène. On dira alors que la liaison entre $X$ et $Y$ fait apparaître les *ordres sous-jacents* $\omega$ et $\omega'$.

Une condition équivalente à l'existence d'ordres sous-jacents à la liaison entre $X$ et $Y$ est l'existence d'un ordre $\omega$ sur l'ensemble de modalités de l'une des variables, disons $E$, tel que pour tout couple $(y,y')$ de modalités de l'autre variable, l'ordre de distorsion $(P_{y'}/P_y)$ entre les distributions conditionnelles de la première variable coïncide soit avec $\omega$ soit avec son dual. Un ordre $\omega'$ s'en déduit alors sur l'ensemble $F$ des modalités de cette seconde variable : c'est l'ensemble des couples $(y,y')$ tels que $(P_{y'}/P_y)$ coïncide avec $\omega$ (on vérifie facilement que cette relation est transitive).

Quand la liaison entre deux variables ne révèle pas deux ordres sous-jacents, on peut utilement rechercher les ordres $\omega$ sur $E$ et $\omega'$ sur $F$ qui sont les plus proches de réaliser cette condition, par exemple ceux qui maximisent l'indice d'homogénéité.

## 7. ANALYSE DE LA CORRÉLATION ENTRE UNE VARIABLE ORDONNÉE ET UNE VARIABLE NON ORDONNÉE

Si l'une des variables, disons $X$, est ordonnée et pas l'autre, on peut rechercher s'il existe un ordre $\omega'$ sur $F$ qui rende la liaison entre les deux variables ordonnées homogène. S'il existe un tel ordre$\omega'$, on dira que c'est *l'ordre induit sur F* par $\omega$. Sinon, on pourra chercher l'ordre le plus proche de réaliser cette condition, toujours en maximisant l'indice d'homogénéité résultant. La recherche de l'ordre induit se fait en calculant les ordres de distorsion $(P_{x'}/P_x)$ pour $(x,x')$ dans $\omega$ : en effet tous ces ordres doivent coïncider avec lui s'il existe. Là encore, il suffit en fait de considérer ces ordres de distorsion pour les couples de modalités $x$ et $x'$ voisines.

L'existence d'un ordre sur $F$ induit par un ordre a priori $\omega$ sur $E$ implique que la liaison entre les deux variables dépourvues d'ordre fasse apparaître des ordres sous-jacents, et que l'ordre sous-jacent sur $E$ soit $\omega$.

Il arrive également que l'analyse non-ordonnée de variables ordonnées fasse apparaître des ordres sous-jacents différents des ordres a priori. Par exemple, l'ordre a priori $\omega$ sur $E$ peut induire sur $F$ un ordre $\omega''$ différent de l'ordre a priori $\omega'$. Les rapports entre $\omega'$ et $\omega''$ peuvent alors être révélateurs d'une liaison non monotone. On rencontre en particulier des situations où $\omega''$ est Blackien par rapport à $\omega'$. C'est typiquement le cas en sciences sociales ou en biologie, dans des phénomènes de sénescence ou de néoténie : $X$ est par exemple une échelle d'opinion ou une mesure biométrique, et $Y$ est l'âge des individus. Si l'on analyse la liaison sans tenir compte de la structure d'ordre naturelle $\omega'$ sur $Y$, il arrive que la structure d'ordre de la variable $X$ induise sur $Y$ un ordre $\omega''$ qui coïncide avec $\omega'$ pour les classes les plus jeunes, puis avec son dual pour les plus âgées, présentant un rebroussement, une sorte de retour en enfance.

## 8. CONCLUSION ET APPEL A EXEMPLES

L'étude de la corrélation, même limitée à l'analyse bi-variée, apparaît donc comme une notion très riche, en particulier si l'on se donne une grande liberté dans le choix des structures a priori. L'auteur serait très reconnaissant envers les lecteurs qui, intéressés par les techniques exposées ici, lui communiqueraient des exemples de situations réelles dans lesquelles l'une d'elles leur semblerait pertinente. Il se propose, avec leur accord, d'assurer la diffusion de ces exemples et de leur traitement auprès de tous ceux qui auront manifesté leur intérêt pour ces applications.

# BIBLIOGRAPHIE

[1] ARABIE P. et HUBERT L.J., «An overview of combinatorial data analysis», in *Clustering and classification*, P.ARABIE, L.J. HUBERT et G. DE SOETE eds, River Edge, World Scientific Publ., 1996

[2] CALOT G., *Cours de statistique descriptive*, Paris, Dunod, 1965

[3] DROESBEKE J.J., *Eléments de statistique*, Paris, Ellipses, 1988

[4] LEHMANN E.L., «Some concepts of dependence», *Annals of Mathematical Statistics,* 37 (1966), 1137-1153.

[5] LERMAN I.C., «Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles» I, *Mathématiques, Informatique et Sciences humaines*, 118 (1992), 33-52.

[6] PETIT J.L. et TÉROUANNE É., *Résumons-nous, Modèles Mathématiques en Sciences Sociales,* Paris, Ellipses, 1988.

[7] POTTIER P., «Mesure de la liaison entre deux variables qualitiatives : relations entre un coefficient de corrélation généralisé et le Chi-deux», *Revue de Statistique Appliquée*, 42 (1994), 41-61.

[8] ROUANET H., LE ROUX B. et BERT M.C, *Statistiques en sciences humaines : procédures naturelles*, Paris, Dunod, 1987.

[9] TÉROUANNE É., «Distorsion entre deux distributions d'une variable nominale», *Mathématiques, Informatique et Sciences humaines*, 131 (1995), 29-38.

[10] TÉROUANNE É., «Comparaison de tendance centrale par l'analyse de transferts», *Mathématiques, Informatique et Sciences humaines*, 134 (1996), 63-76.