Statistique et analyse des données

G. Le Calve

Un indice de similarité pour des variables de types quelconques

Statistique et analyse des données, tome 1, n° 2 (1976), p. 39-47

<http://www.numdam.org/item?id=SAD_1976__1_2_39_0>

UN INDICE DE SIMILARITE POUR DES VARIABLES DE TYPES QUELCONQUES

G. LE CALVE (x)

Ce qui suit est un condensé de (2). Nous avons insisté surtout sur l'aspect épistémologique ; pour l'aspect mathématique, nous nous sommes contentés d'énoncer quelques résultats, renvoyant le lecteur curieux à (2) où il pourra trouver, outre les démonstrations, des formules détaillées pour les cas classiques ainsi que des exemples d'application.

x) Université de Haute Bretagne

INTRODUCTION

En sciences humaines, la plupart des problèmes que le statisticien a à traiter, se trouvent posés de la manière suivante : un tableau rectangulaire de données numériques ou prétendues telles, croise un ensemble $\mathcal{C}$ de caractères et un ensemble E d'individus. Il s'agit alors d'interpréter ce tableau en en tirant, par exemple, des typologies sur les caractères ou les individus. Pour ce faire, toutes les méthodes remplacent ce tableau de départ par un tableau d'indices de "similarité" ou de "proximité" (covariance par exemple).

Le problème qui se pose alors est que les caractères sont très rarement de même nature et que s'il est possible de calculer une proximité entre deux équivalences ou deux ordres, par exemple, la proximité entre deux questions de nature différente reste à définir.

C'est à ce niveau que se situe notre étude. Notre propos n'est pas de définir un indice venant s'ajouter à la longue liste de ceux déjà connus, mais, au contraire, d'en trouver un qui puisse recouvrir comme cas particulier de nombreux indices et permettre la comparaison de questions de nature différente ou même n'entrant pas dans le cadre habituellement traité.

La statistique que nous définirons n'est autre qu'une généralisation de la statistique due à Lerman. Celle-ci tient parfaitement compte de la nature des questions quand elles sont toutes deux du même type classique mais permet également de définir la vraisemblance de la proximité entre deux variables. Ce dernier aspect nous semble capital, car il est plus facile, et donc chargé de moins d'information, de se rassembler dans des structures pauvres que dans des structures riches.

I - UN INDICE DE SIMILARITE

Reprenons l'exemple du coefficient d'association de K. Pearson, tel qu'on le trouve décrit dans Lerman :

X et Y étant deux sous-ensembles de E, on pose S = Card $(X \cap Y)$. Cette statistique dépendant par trop de la fréquence des attributs X et Y indépendamment de leur liaison, nous normaliserons en prenant $U_0 = \frac{\Delta - \bar{S}}{\sqrt{\mathrm{Var}\, S}}$ où S est la variable aléatoire Card $(A \cap B)$, A (resp. B) parcourant l'ensemble des parties de E de même cardinal que X (resp. Y)

Soient maintenant X et Y deux variables qualitatives sur E, c'est à dire deux relations binaires représentées par leur matrice $X = (x_{ij})$, $Y = (y_{ij})$. Calquant la méthode précédente, nous choisirons pour S le cardinal de l'intersection des graphes de X et Y dans E^2, et considèrerons la v.a. Card $(A \cap B)$ où A (resp.B) est un sous-ensemble de E^2 de même cardinal que X (resp.Y). Un calcul simple donne :

$$U_0 = \frac{\left(\sum_{ij} x_{ij} y_{ij} - \frac{1}{n^2}\bar{X}\bar{Y}\right)\sqrt{n^2-1}}{\sqrt{\sum_{ij}\left(x_{ij} - \frac{\bar{X}}{n^2}\right)^2 \sum_{ij}\left(y_{ij} - \frac{\bar{Y}}{n^2}\right)^2}}$$

en posant $\sum_{ij} x_{ij} = \bar{X}$, $\sum_{ij} y_{ij} = \bar{Y}$

On peut remarquer que si $S = \sum_{ij} x_{ij} y_{ij}$ est le produit xalaire des matrices X et Y, $\frac{u_0}{\sqrt{m^2-1}}$ en est alors le coefficient de corrélation.
Mais cet indice s'avère mauvais.

En effet si u_0 tient bien compte du nombre de 1 des matrices X et Y, il ne tient absolument pas compte de leur disposition, c'est à dire de la nature des matrices.

La solution semble donc de faire parcourir à A et B l'ensemble des matrices de même type que X et Y respectivement, le problème étant de définir ce qu'on appelle le type d'une matrice.

Lorsqu'un expérimentateur prépare une question X à poser à une population, il prépare en même temps la codification des réponses, par exemple une équivalence à 3 classes notées 0, 1, 2. La tentation est alors grande de considérer le type de la question comme étant une équivalence à 3 classes. C'est supposer que le type de la question est fonction uniquement de la connaissance que l'expérimentation a, à priori, du problème, et pas du tout de la population à laquelle elle s'applique (il peut avoir introduit des classes vides). Nous tiendrons donc compte de l'effectif de chacune des classes et dirons que le type de la matrice est une équivalence à 3 classes d'effectifs respectifs fixés n_0, n_1, n_2 . Ceci donne la définition suivante :

DEFINITION

Deux matrices X et A seront dites de même type si elles se déduisent l'une de l'autre par une permutation; autrement dit $X = \sigma A \sigma^t$ *où* σ *est une matrice de permutation sur E.*

LA VARIABLE ALEATOIRE $S_{X,Y}$

Soit E un ensemble de n individus. Pour toute question X on définit une fonction $(i,j) \to x_{ij}$ de E^2 dans IR vérifiant $x_{ii} = 0 \forall i$. Cette fonction est parfois appelée un score. La convention $x_{ii} = 0$ est utilisée uniquement pour se débarasser du problème de la réflexivité.

On définit $S_{X,Y} = \sum_{ij} x_{ij} y_{ij} = Tr(X Y^t)$

Soit $\mathfrak{S}(E)$ l'ensemble des matières de permutation sur E.

Posons $\Omega = \mathfrak{S}(E) \times \mathfrak{S}(E)$ et munissons Ω d'une mesure de probabilité uniformément répartie. A tout couple de caractères X, Y on associe la variable aléatoire $S_{X,Y}$ définie par $S_{X,Y}(\sigma,\sigma') = Tr(\sigma X \sigma'^t Y^t \sigma')$

Nous désignerons par $U_{X,Y}$ la variable centrée réduite

$$U_{X,Y}(\sigma,\sigma') = \frac{S_{X,Y}(\sigma,\sigma') - \bar{S}}{\sqrt{\mathrm{Var}\, S_{X,Y}}} \qquad U^c_{X,Y} = \frac{S_{XY} - \overline{S_{XY}}}{\sqrt{\mathrm{Var}\, S_{XY}}}$$

./...

Nous désignerons enfin par $P_{x,y}$ la probabilité que $U_{x,y}$ soit inférieure à $U^{o}_{x,y}$.

Comme le note Lerman "$P_{x,y}$ définit une mesure de la ressemblance entre les variables où la notion de ressemblance est clairement remplacée par la notion de "vraisemblance" ".

II - ETUDE DE LA VARIABLE ALEATOIRE $S_{x,y}$

Lerman a démontré que lorsque X et Y sont simultanément des attributs ou des équivalences ou des préordres, il revient au même de fixer un quelconque des caractères et de choisir un autre au hasard parmi les caractères du même type que le second.

La forme matricielle utilisée içi permet de démontrer facilement le

THEOREME DE DUALITE

Quels que soient les caractères X et Y on a

$$S_{x,y}(\sigma,\sigma') = S_{x,y}(\sigma'\circ\sigma, I) = S_{x,y}(I, \sigma'\circ\sigma) \quad \forall\sigma, \forall\sigma'$$

La démonstration, immédiate, résulte de la définition de $S_{x,y}(\sigma,\sigma')$ *et du fait que Tr (AB) = Tr (BA).*

CARACTERISTIQUES DE $S_{x,y}$

Le théorème précédent apporte une simplification des calculs de la moyenne et de la variance de $S_{x,y}$.

On trouve $\overline{S}_{x,y} = \dfrac{\overline{X}\,\overline{Y}}{n(n-1)}$ où $\overline{X} = \sum_{ij} x_{ij}$

Le calcul de la variance fait intervenir des expressions du genre $\sum_{ijkl} x_{ij}\,x_{kl}$ Nous les noterons V_4X, V_3X, V_2X suivant que la sommation s'effectue sur 4, 3 ou 2 indices distincts. De plus, nous décomposerons V_3X en quatre et V_2X en deux sommes partielles, suivant les positions occupées par les indices distincts. On a alors :

$$\mathrm{Var}\, S_{x,y} = \frac{V_4X\,V_4Y}{n(n-1)(n-2)(n-3)} + \frac{V_3'X\,V_3'Y + V_3''X\,V_3''Y + V_3'''X\,V_3'''Y + V_3''''X\,V_3''''Y}{n(n-1)(n-2)} + \frac{V_2'X\,V_2'Y + V_2''X\,V_2''Y}{n(n-1)} - \overline{S}_{x,y}^{\,2}$$

Cette formule se simplifie considérablement dans les cas où X et Y sont des variables classiques (cf. [2] où l'on pourra trouver les formules explicites).

./...

Tendance vers la loi normale

On démontre d'abord le

Lemme

Soient A et B deux matrices carrées $u \times u$ telles que $\sum_{ij} a_{ij} = \sum b_{ij} = 0$

Soit Ω l'ensemble des matrices n x n de permutations. Alors la distribution de la variable aléatoire $T(\sigma) = T_n(A^t B^t \sigma)$ tend vers la loi normale quand n tend vers l'infini si l'hypothèse H est vérifiée.

Hypothèse H

Les quantités a_{ij} et b_{ij} sont uniformément bornées et $\sum_{ijl} a_{ij} a_{il}$, ainsi que les quantités qui s'en déduisent par permutation sur les indices i, j, l ou échange de a en b, sont d'ordre n^3

On en déduit le

Théorème

Désignons les modalités de X (resp.Y) par a (resp.α) et leur effectif par n_a(resp n_α). Si pour tout a et tout α les rapports $\frac{n_a}{n}$ et $\frac{n_\alpha}{n}$ tendent vers des limites lorsque n tend vers l'infini, alors la distribution de probabilité de $U_{x,y}$ tend vers la loi normale centrée réduite.

Pour la démonstration, nous renvoyons à [2]

III - QUELQUES CAS PARTICULIERS

III - 1. Quelques indices connus

Les caractères X et Y ne prennent que les valeurs 1 et 0 interprétées comme présence-absence.

Nous définirons les scores par x_{ij} = 1 si i possède l'attribut X, x_{ij} = 0 autrement(*) Toutes les colonnes de la matrice X (resp.Y) sont donc égales. Si le tableau de contingence de X et Y est

	Y	$\bar{Y}$
X	a	u
$\bar{X}$	v	c

on trouve

$$U^0_{x,y} = \frac{\sqrt{n}\,(at - uv)}{\sqrt{(a+u)(a+v)(t+u)(t+v)}}$$

./...

(*) et $x_{ii} = 0$

C'est-à-dire le coefficient de K. Pearson, à une constante multiplicative près

- X et Y induisent un ordre total sur E.

Posons $x_{ij} = 1$ si i est avant j pour l'ordre induit par X et la même convention pour y_{ij}. Alors $U^0_{x,y}$ n'est autre que le τ de Kendall.

Si l'on pose x_{ij} = rang de i, on trouve alors le coefficient de Spearman, à une constante multiplicative près

- X et Y sont des variables réelles

En posant $x_{ii} = 0$, $x_{ij} = X(i)$ la valeur de la variable pour i, on obtient pour $S_{x,y}$ le produit scolaire et pour $U^0_{x,y}$ le coefficient de corrélation. à une constante multiplicative près.

- critères de classificabilité

Soient D un ensemble de variables à classifier, U la matrice des proximités sur D, Θ une relation d'équivalence sur D et K l'ordre sur D^2 engendré par U.

Les quantités $\Delta_{U,\Theta}$ ne sont autres que les critères de classificabilité étudiés en détail par Lerman dans [1]

III - 2. quelques cas nouveaux

- La codification 1 - 0 représentant détection - non détection.

La codification 1 - 0 peut être interprétée non comme présence-absence, mais comme détection non détection. Dans cette optique, seule la codification 1 apporte de l'information. Il est clair que le coefficient de Pearson, invariant par changement simultané du codage de X et de Y ne peut répondre au problème.

Nous posons $x_{ij} = 1$ si i et j ont l'attribut X, 0 autrement. Ceci conduit à représenter X par une matrice de la forme

1	0

et donne des résultats différents du coefficient de Pearson.

- Préodres non totaux

Parmi les cas non classiques le préordre non total est certainement le plus répandu (par exemple : niveaux de formation scolaire, ou catégorie socio-professionnelles).
Il est en général, traité comme une relation d'équivalence et il y a donc perte d'information. L'indice décrit ici permet de prendre en compte le préordre non total avec sa structure particulière.

- Variables hétérogènes

Dans tous les indices classiques, il faut supposer que les variables à comparer sont de même nature. Ceci se fait, en général, en appauvissant les structures riches pour toutes les ramener au même niveau, ce qui entraîne une perte considérable d'information, ou par une technique mixte d'appauvissement de certaines structures et d'enrichissement d'autres, ce qui entraîne un biais.

./...

L'indice décrit içi n'a pas ces travers puisque, manifestement, rien n'oblige X et Y a être de même nature, répondant par là à l'objectif principal que nous nous étions fixé.

- Données sous formes matricielles

Nous n'avons jusqu'à présent décrit que des situations où la réponse de la population à la question X se présentait sous forme d'un vecteur de réponse, vecteur transformé en une matrice à l'aide de la fonction score. Dans de nombreux problèmes, les données sont directement sous forme matricielle : comparaison par paires, matrices de transport entre différents points, de communication, de domination, d'incidences, etc...

IV - QUELQUES PROBLEMES LIES AUX QUESTIONNAIRES

Ces différents problèmes ne devraient pas se poser si les questionnaires étaient construits et remplis de manière parfaite. Mais quelque soit le soin que l'on puisse apportér à ces deux points, l'imperfection de la nature humaine veut que ces problèmes subsistent.

IV - 1. Données manquantes

Nous n'oserions prétendre résoudre ce problème épineux et important en quelques lignes, mais apporter des indications sur une conduite à tenir qui serait conforme à l'esprit de la méthode.

Nous distinguerons deux cas suivant que l'absence de donnée est porteuse d'information ou non.

La donnée peut être absente pour des raisons purement matérielles : elle n'a pas été posée à l'individu i, sa réponse a été égarée, ou était illisible ou mal codée etc... Cette absence de donnée ne porte aucune information et nous retirerons l'individu i de la population pour la question X, $S_{x,y}$ sera calculée en tenant compte de la population ayant répondu aux deux questions.

La donnée peut être absente parce que l'individu i a refusé de répondre à la question X, ne se reconnaissant dans aucune des modalités offertes. Ce cas est tout différent du précédent et il ne convient pas de la traiter de la même manière. Revenant à la définition de la matrice (X_{ij}) nous écrirons que
(i ne peut être mis en relation avec quelque individu que ce soit).

La différence entre les deux méthodes n'est pas au niveau du calcul de $S_{x,y}$ qui est le même dans les deux cas, mais au niveau des moyennes et des variances qui sont différentes. Si, par exemple, deux variables X et Y sont considérées comme identiques par les rares individus qui ont accepté d'y répondre, leur indice $U_{x,y}$ de proximité sera bien plus grand dans la seconde technique que dans la première.

IV - 2. Réponses multiples

Un autre cas presque aussi fréquent que le précédent est celui çi : l'individu i coche deux réponses au lieu d'une seule. Là encore, il faut essayer de distinguer si ce double choix implique une adhésion aux deux opinions ou une hésitation entre les deux. En désignant par x'_{ij} (resp x''_{ij}) ce qu'aurait été x_{ij} si i avait opté pour la première opinion (resp. la seconde, nous poserons $x_{ij} = Max(x'_{ij}, x''_{ij})$ dans le cas de la double adhésion et $x_{ij} = \frac{1}{2} x'_{ij} + \frac{1}{2} x''_{ij}$ dans le cas d'hésitation.

Dans certains questionnaires, les réponses multiples sont prévues et organisées dès le départ puisque, au lieu d'une adhésion à une opinion, on demande aux sujets de donner une note aux différentes opinions. En désignant par α_i^S la note donnée par i à l'opinion S, par x_{ij}^{ST} ce qu'aurait été x_{ij} si i avait opté pour S et J pour T

On posera $$x_{ij} = \frac{\sum_{S,T} \alpha_i^S \alpha_j^T x_{ij}^{ST}}{\sum_{S,T} \alpha_i^S \alpha_j^T}$$

IV - 3. Crédibilité des données

En Sciences Humaines, l'expérimentateur est rarement maître totalement des données qu'il manipule et il lui est, en général, impossible de recommencer telle mesure qui lui semble suspecte. Bien souvent il sait que telle source de renseignements est sujette à caution (cas très fréquent en histoire par exemple). L'expérimentateur peut affecter à chacun des individus un coefficient β_i de crédibilité (dans la plupart des cas que nous avons pu rencontrer ce coefficient ne prend que 2 ou 3 valeurs), coefficient compris entre 0 et 1. En désignant par (x'_{ij}) ce qu'aurait été la matrice en l'absence de ces coefficients de crédibilité, nous poserons $x_{ij} = \beta_i \beta_j x'_{ij}$

Cette notion de crédibilité des données peut évidemment se combiner avec les réponses multiples et/ou les données absentes. Dans un cas comme dans l'autre, cela revient à considérer X comme une relation floue sur E.

V - LES VARIABLES SONT DES RELATIONS DE E DANS F.

Nous considèrerons également la variable aléatoire obtenue en tirant au hasard X et Y parmi les matrices de $\hat{\omega}$-type, en prenant la définition suivante.

On peut également avoir à traiter des données se présentant sous forme de matrices rectangulaires : tableaux d'incidences, matrices de communications, réseaux de transports, etc...

La méthode s'étend très facilement à ce cas. Nous prendrons comme base de la statistique $S_{x,y} = tr\,(X, Y^t)$.

DEFINITION

Soient X_1 et X_2 deux matrices sur E × F. Elles seront dites de même type si il existe deux matrices de permutation A sur E et B sur F telles que $X_1 = A\, X_2\, B$.

On a également le

THEOREME DE DUALITE

Les 3 variables aléatoires $S^{i}_{X,Y}(\cdot,\cdot)$ $i \leq 3$ *obtenues en*

- *tirant au hasard deux matrices parmi les matrices de même type que X et de même type que Y,*
- *fixant X et tirant au hasard une matrice de même type que Y,*
- *fixant Y et tirant au hasard une matrice de même type que X*

ont la même loi de probabilité.

En effet, en désignant par A et A' (resp. B et B') des matrices de permutation sur E (resp. sur F)

Les 3 variables aléatoires $S^{i}_{X,Y}(\cdot,\cdot)$ $i \leq 3$ *obtenues en*

- *tirant au hasard deux matrices parmi les matrices de même type que X et de même type que Y,*
- *fixant X et tirant au hasard une matrice de même type que Y,*
- *fixant Y et tirant au hasard une matrice de même type que X*

ont la même loi de probabilité.

En effet, en désignant par A et A' (resp. B et B') des matrices de permutation sur E (resp. sur F).

$$S_{x,y}(A,B,A',B') = Tr\,(A\,X\,B\,B'^{t}\,Y^{t}\,A'^{t})$$
$$= Tr\,(A'^{t}\,AXB\,B'^{t}\,Y^{t}) = S_{x,y}(A'^{t}oA, BoB'^{t}, I, I)$$
$$= Tr\,(XBB'^{t}\,Y^{t}A'^{t}A) = S_{x,y}(I,I,A^{t}oA', B'oB^{t})$$

Caractéristiques de la Va $S^{(w)}_{x,y}$

Nous désignerons par n le nombre d'éléments de E et par p le nombre d'éléments de F et par a et a' (resp b et b') des permutations sur E (resp sur F).

Avec des notations analogues au cas des matrices $n \times n$ on obtient :

$$\bar{S}_{x,y} = \frac{\bar{X}\bar{Y}}{np}$$

$$Var\, S_{x,y} = \frac{V_4 X V_4 Y}{n(n-1)p(p-1)} + \frac{V_3 X V_3 Y}{np(p-1)} + \frac{V'_3 X V'_3 Y}{n(n-1)p} + \frac{V_2 X V_2 Y}{np} - \frac{\bar{X}^2\bar{Y}^2}{n^2p^2}$$

B I B L I O G R A P H I E

1 I.C. LERMAN. "Etude distributionnelle de statistiques de proximité entre structures algébriques finies du même type; application à la classification automatique" Cahiers du B.U.R.O. 1972 - N° 19.

2 G. LE CALVE. "Un indice de similarité pour des variables de type quelcon-