MATHÉMATIQUES ET SCIENCES HUMAINES

YVES LIGNON

Corrélation entre deux variables ordinales dont l'une est dichotomisée

Mathématiques et sciences humaines, tome 76 (1981), p. 47-57

<http://www.numdam.org/item?id=MSH_1981__76__47_0>

Math. Sci. hum. (19e année, n°76, *1981*, p.47-57)

CORRELATION ENTRE DEUX VARIABLES ORDINALES DONT L'UNE EST DICHOTOMISEE

Yves LIGNON*

On expose une méthode permettant de mesurer la corrélation entre deux échantillons de valeurs de variables ordinales X et Y, X étant dichotomisée. On étudie le cas où l'échelle Y comporte des ex-æquo. Ensuite, après avoir constaté que le problème étudié est équivalent à celui de Wilcoxon-Mann-Whitney, on montre que le coefficient obtenu peut être considéré comme la forme particulière de coefficients classiques : τ de Kendall, ρ de Spearman et (dans quelques cas) bisérial.

I. POSITION DU PROBLEME

I.1. Rappel : le τ de Kendall. On sait que la corrélation entre deux échantillons de valeurs de variables ordinales peut être mesurée au moyen des coefficients ρ (Spearman) ou τ (Kendall).

Dans le cas de l'absence d'ex-æquo, Kendall (1938-1975) définit τ comme suit : soient N sujets tels que le classement successif sur deux échelles ordinales permette d'associer au sujet i le couple (x_i, y_i) [1] ; la paire de sujets (i,j) est dite accordée :

si $x_i < x_j$ et $y_i < y_j$ ou si $x_i > x_j$ et $y_i > y_j$

et désaccordée :

si $x_i < x_j$ et $y_i > y_j$ ou si $x_i > x_j$ et $y_i < y_j$

alors :
$$\tau = \frac{S_+ - S_-}{S_{max}} \qquad (1.1.)$$

---

* UER Mathématiques - Université de Toulouse-le-Mirail.

(1) Kendall (1975) note (paragraphe 1.2.) "it is customary but not essential to denote the ranks by ordinal numbers".

avec S_+ : nombre de paires accordées, S_- : nombre de paires désaccordées,

S_{max} : $\frac{1}{2}$ N(N-1) nombre total de paires.

On vérifie facilement que τ est un coefficient de corrélation dont la distribution, ou plutôt celle de $S = S_+ - S_-$ est connue.

I.2. Le cas des ex-æquo.

Si X ou Y ou les deux comportent des ex-æquo il est impossible de dire de certaines paires si elles sont accordées ou non au sens de I.1. Kendall (1945) a proposé, après avoir attribué aux ex-æquo la moyenne des places qu'ils occuperaient s'ils ne l'étaient pas, d'exclure de telles paires de l'analyse et de poser :

$$\tau = \frac{S_+ - S_-}{S_{max\text{-}corrigé}} \qquad (1.2.)$$

avec : $$S_{max\text{-}corrigé} = \left\{\left[\frac{1}{2} N(N-1) - T\right]\left[\frac{1}{2} N(N-1) - U\right]\right\}^{1/2} \qquad (1.3.)$$

où : $$T = \frac{1}{2} \Sigma\, t_i(t_i-1) \qquad U = \frac{1}{2} \Sigma\, u_i(u_i-1)$$

t_i (resp. u_i) étant le nombre de sujets constituant le $i^{ème}$ ensemble d'ex-æquo du classement X (resp. Y).

Whitfield (1947) et Kendall (1948-1975) ont étendu le coefficient (1.2.) au cas où X est dichotomisée en considérant qu'alors l'échantillon des x_i est composé de deux ensembles d'ex-æquo. Un contre-exemple ci-dessous mettra en évidence ce qui nous semble être un inconvénient de ce coefficient.

I.3. Corrélation maximum dans le cas où X est dichotomisée. Conditions de Cureton.

I.3.1. Corrélation maximum. Soit n_0 (resp. n_1) l'effectif de la valeur 0 (resp. 1) de X. Un sujet classé 0 se situant "avant" un sujet classé 1, on peut considérer qu'il y a corrélation maximum positive (resp. négative) lorsque les n_0 sujets classés 0 sur l'échelle X occupent les places 1 à n_0 sur l'échelle Y (resp. les n_0 sujets classés 0 sur l'échelle X occupent les places n_1+1 à N sur l'échelle Y).

I.3.2. Les conditions de Cureton. Il résulte de ce qui précède que, pour pouvoir être considérée comme un coefficient de corrélation au sens de Kendall, la mesure de la liaison statistique entre deux échantillons de valeurs de variables ordinales X et Y, X étant dichotomisée doit satisfaire aux conditions suivantes posées par Cureton (1956).

a) la mesure prend la valeur 1 lorsque les n_1 sujets classés 1 sur l'échelle X occupent les n_1 places les plus élevées sur l'échelle Y.
b) la mesure prend la valeur -1 lorsque les n_0 sujets classés 0 sur l'échelle X occupent les n_0 places les plus élevées sur l'échelle Y.
c) la mesure est définie uniquement en termes de paires accordées ou désaccordées. Ce faisant elle est non paramétrique.

On notera que ces conditions ne font appel à aucun aspect particulier de l'échelle Y et restent donc valables en présence d'ex-æquo.

I.3.3. Un contre-exemple. Soient les deux échantillons :

Sujets	X	Y
A	0	1
B	0	2
C	0	3
D	1	4
E	1	5

$N = 5$

$n_0 = 3,\ n_1 = 2$

$S_+ = 6 \quad S_- = 0$

$t_1 = 3 \quad t_2 = 2$

$u_i = 0$

le coefficient de Whitfield-Kendall vaut 0,77 alors que la corrélation est maximum positive au sens de Cureton et que d'ailleurs toutes les paires retenues pour l'analyse sont accordées.

II. LE COEFFICIENT RANG-BISERIAL. CAS OU IL N'Y A PAS D'EX-ÆQUO SUR L'ECHELLE Y.

II.1. Définition

Soit τ_{bs} le coefficient de corrélation cherché. Nous posons :

$$\tau_{bs} = \frac{S_+ - S_-}{D} \qquad (2.1.)$$

où S_+ et S_- ont leurs sens habituels et où D est tel que :

$$S_+ = 0 \Rightarrow \tau_{bs} = -1$$

$$S_- = 0 \Rightarrow \tau_{bs} = +1$$

II.2. Paires accordées et désaccordées

La paire de sujets (i,j) est dite :

accordée si et seulement si : $x_i = 0 ; x_j = 1$ et $y_i < y_j$ ou

$x_i = 1 ; x_j = 0$ et $y_i > y_j$

désaccordée si et seulement si : $x_i = 0 ; x_j = 1$ et $y_i > y_j$ ou

$x_i = 1 ; x_j = 0$ et $y_i < y_j$

Cette extension au cas particulier traité de la définition des paires accordées ou désaccordées conduit à exclure les paires telles que $x_i = x_j$ qui sont celles exclues par la méthode de Whitfield-Kendall. Il s'ensuit que le numérateur de τ_{bs} est le même que celui du coefficient de Whitfield-Kendall.

II.3. Définition de D

Utilisant l'un des points de vue à partir desquels Kendall (1948-1975) définit τ nous prenons pour valeur de D le nombre de paires figurant dans l'analyse soit :

$$D = S_+ + S_- = n_0 n_1 \qquad (2.2.)$$

puisque chaque sujet classé 0 en X est soit accordé, soit désaccordé avec chacun des n_1 sujets classés 1 et que les paires ainsi constituées sont les seules conservées, dès lors :

$$\tau_{bs} = \frac{S_+ - S_-}{n_0 n_1} \qquad (2.3.)$$

expression à partir de laquelle on vérifie facilement que les conditions de Cureton sont satisfaites (on constatera qu'on a bien $\tau_{bs} = +1$ pour les données de I.3.3.). Dans son article de 1956 cet auteur avait introduit τ_{bs} sur un exemple à partir duquel il avait vérifié la satisfaction des conditions posées au départ.

II.4. Méthode pratique de calcul. Exemple.

On peut adapter la méthode classique de calcul de τ en partant d'un tableau à 3 colonnes : colonne 1 : sujets ; colonne 2 : valeurs de X ; colonne 3 : valeurs de Y écrites de haut en bas dans l'ordre croissant.

Avec cette disposition :

- tout sujet classé 0 en X est accordé avec tout sujet classé 1 et situé "en dessous de lui" dans le tableau.
- tout sujet classé 1 en X est désaccordé avec tout sujet classé 0 et situé "en dessous de lui" dans le tableau. Donc :

avec v_i : nombre de 1 situés sous le $i^{ème}$ 0 de la colonne 2

w_j : nombre de 0 situés sous le $j^{ème}$ 1 de la colonne 3

$$S_+ = \Sigma v_i \qquad S_- = \Sigma w_j$$

Ainsi soit le tableau (pour lequel Kendall (1948-1975) trouvait $\tau = 0{,}24$).

Sujets	X	Y
A	0	1
B	0	2
C	1	3
D	0	4
E	1	5
F	1	6
G	0	7
H	0	8
I	0	9
J	1	10
K	0	11
L	1	12
M	0	13
N	1	14
O	1	15

$N = 15$

$n_0 = 8 \quad n_1 = 7$

$D = 56$

$S_+ = 7+7+6+4+4+4+3+2 = 37$

$S_- = 6+5+5+2+1 = 19$

$$\tau_{bs} = \frac{37 - 19}{56} = 0,32$$

III. CAS OU IL Y A DES EX-ÆQUO SUR L'ECHELLE Y

III.1. Définitions

Ce sont les même qu'en II. Dès lors sont exclues les paires telles que l'une au moins des deux égalités $x_i = x_j$; $y_i = y_j$ soit vraie.

III.2. Calcul de D

Nous proposons ici une solution générale de ce problème résolu dans un cas particulier par Cureton (1956).

D étant toujours le nombre de paires figurant dans l'analyse, les conditions posées en I.3.2. sont "ipso-facto" satisfaites. Le problème réside dans la détermination explicite de la valeur de D.

Les paires exclues sont celles, telles que :

a) $x_i = x_j$ $\forall$ y_i et y_j , en nombre s_1

b) $x_i \neq x_j$ et $y_i = y_j$, en nombre s_2

on a vu en II. que $\frac{1}{2} N(N-1) - s_1 = n_0 n_1$

donc ici : $D = n_o n_1 - s_2$ (3.1.)

Etendant l'appellation choisie par Cureton nous désignerons par "bracket-tie" un ensemble de sujets ex-æquo en Y mais non tous classés au même niveau en X et soit alors dans le $k^{ème}$ bracket-tie l_k (resp. m_k) le

nombre de sujets tels que $x_i = 0$ (resp. $x_i = 1$). On peut donc constituer à partir de ce bracket-tie $l_k\ m_k$ paires qui doivent être exclues en raison de b), il s'ensuit que :

$$s_2 = \Sigma\ l_k\ m_k \qquad \text{et que}$$

$$D = n_0 n_1 - \Sigma\ l_k\ m_k \qquad \text{soit}$$

$$\tau_{bs} = \frac{S_+ - S_-}{n_0 n_1 - \Sigma\ l_k\ m_k} \qquad (3.2.)$$

Cureton (1956) avait fourni une solution simplifiée relative au cas d'un seul bracket-tie (ce cas se présente entre autres lorsque la corrélation est maximum).

III.3. Application numérique

III.3.1. Méthode de calcul : avec la même disposition des données qu'en II.4.

- tout sujet classé 0 en X est accordé avec tout sujet classé 1 et situé "au dessous de lui" *exclusion faite des sujets classés au même rang que lui en Y.*
- tout sujet classé 1 en X est désaccordé avec tout sujet classé 0 et situé "au dessous de lui" *exclusion faite des sujets classés au même rang que lui en Y.*

avec v'_i : nombre de 1 situés sous le $i^{ème}$ 0 de la colonne 2 et correspondant à des valeurs de Y différentes de y_i

w'_j : nombre de 0 situés sous le $j^{ème}$ 1 de la colonne 2 et correspondant à des valeurs de Y différentes de y_j

$$S_+ = \Sigma\ v'_i \qquad S_- = \Sigma\ w'_j$$

III.3.2. Exemples

a) données proposées par Cureton (1956)

Sujets	X	Y
A	0	1
B	0	2,5
C	0	2,5
D	1	4,5
E	1	4,5
F	0	6,5
G	1	6,5
H	1	8
I	1	9,5
J	1	9,5

$N = 10$

$n_0 = 4 \quad n_1 = 6$

"bracket-tie" : (F,G)

$l_1 = 1 \quad m_1 = 1$

$S_+ = 6+6+6+3 = 21$

$S_- = 1+1 = 2$

$$\tau_{bs} = \frac{21 - 2}{24 - 1} = 0,826$$

résultat trouvé par Cureton à la suite de l'étude des 45 paires.

b)

Sujets	X	Y
A	0	1
B	1	3
C	0	3
D	1	3
E	0	5,5
F	1	5,5
G	1	7

$N = 7$

$n_0 = 3 \quad n_1 = 4$

"bracket-ties" (B,C,D)

$l_1 = 1 \quad m_1 = 2$

(E,F) :

$l_2 = 1 \quad m_2 = 1$

$$S_+ = 4 + 2 + 1 = 7$$

$$S_- = 1 + 1 \quad = 2$$

$$\tau_{bs} = \frac{7 - 2}{12 - 3} = 0,555$$

La définition même de τ_{bs} implique que la place relative des sujets d'un même "bracket-tie" à l'intérieur de celui-ci n'a aucune importance et ne modifie pas la valeur du coefficient.

IV. RELATIONS ENTRE τ_{bs} ET D'AUTRES STATISTIQUES

IV.1. τ_{bs} et le problème de Wilcoxon-Mann-Whitney.

Si l'on considère que l'ensemble des sujets peut, au moyen de l'échelle X être subdivisé en "sous-ensemble 0" et "sous-ensemble 1" le problème de l'existence d'une corrélation significative entre X et Y équivaut au problème de l'existence d'une différence significative au sens de Wilcoxon-Mann-Whitney.

L'hypothèse nulle de l'indépendance entre X et Y est en effet l'hypothèse de la même distribution selon Y du "sous-ensemble 0" et du "sous-ensemble 1", l'hypothèse alternative de l'existence d'une corrélation significative est l'hypothèse selon laquelle, sur l'échelle Y, l'un des deux sous-ensembles est statistiquement "plus élevé" que l'autre.

La formulation précédente est exactement celle que, dans son ouvrage classique Siegel (1956) utilise pour présenter le test de Mann et Whitney.

Les quantités S_+ et S_- n'étant d'ailleurs autres que les quantités U et U' qui interviennent dans le dit test de Mann et Whitney, il est dès lors trivial de constater que l'hypothèse nulle selon laquelle τ_{bs} diffère de 0 uniquement à cause des fluctuations d'échantillonnage peut être éprouvée au moyen de ce test.

IV.2. Relations entre τ_{bs} et d'autres coefficients de corrélation.

IV.2.1. Avec le τ de Kendall et le ρ de Spearman.

Les conditions de Cureton font que τ_{bs} peut être considéré comme un coefficient de corrélation de rangs de Kendall mais il existe aussi un lien entre τ_{bs} et le ρ de Spearman.

Durbin et Stuart (1951) ont montré qu'on peut écrire :

$$\rho = \frac{(U - V)}{(U - V)_{max}} \qquad (4.1.)$$

où U (resp. V) est la somme pondérée des paires accordées (resp. désaccordées) la pondération étant obtenue, pour chaque paire, en effectuant la différence des places occupées sur l'une des deux échelles. Tiré d'une idée de Kendall, ce résultat permet d'ailleurs de retrouver certaines des propriétés de ρ énoncées par cet auteur (voir Kendall (1948-1975) chapitres 2 et 3).

Dans notre cas nous pouvons (d'après une autre idée de Kendall) attribuer à chacun des n_0 sujets classés 0 sur l'échelle X, la place $\frac{n_0 + 1}{2}$ et à chacun des n_1 sujets classés 1 la place $n_0 + \frac{n_1 + 1}{2}$ dès lors si nous pondérons toute paire à partir de la différence des places occupées sur X, toutes les pondérations sont égales à $\frac{n_0 + n_1}{2}$ et il s'ensuit que τ_{bs} est une forme particulière de (4.1.).

L'existence de cette propriété avait déjà été constatée par Cureton (1956).

IV.2.2. Avec le coefficient bisérial r_b

IV.2.2.1. Le coefficient de Glass-Cureton.

Glass (1966) et Cureton (1968) ont tiré des résultats de Brogden (1949) la forme particulière du coefficient bisérial lorsque X est ordinale dichotomisée et Y ordinale, l'échantillon des y_i pouvant comporter des ex-æquo.

Alors (Cureton) :

$$r_b = \frac{\overline{Y}_1 - \overline{Y}}{\overline{Y}' - \overline{Y}} \qquad (4.2.)$$

$\overline{Y} = \frac{N + 1}{2}$ moyenne des y_i

$\overline{Y}_1$: moyenne (en Y) des sujets du "sous-ensemble 1"

$\overline{Y}'$: moyenne des n_1 plus grandes valeurs de Y.

IV.2.2.2. Cas sans ex-æquo

alors : $$\overline{Y}' = n_0 + \frac{n_1 + 1}{2}$$

donc : $$r_b = \frac{2}{n_0}\left[\overline{Y}_1 - \frac{N + 1}{2}\right] \qquad (4.3.)$$

Les données étant disposées comme pour le calcul de τ_{bs} soit y_{i1} la place occupée en Y par le $i^{ème}$ sujet classé 1 en X. Le nombre θ_i de 0 situés "au dessus" de ce sujet est tel que :

$$y_{i1} = \theta_i + i$$

et si λ_i est le nombre de paires désaccordées dans lesquelles figure ce sujet

$$\lambda_i = n_0 - \theta_i = n_0 - (y_{i1} - i)$$

d'où : $$S_- = \Sigma \lambda_i = n_0 n_1 - \Sigma y_{i1} = \frac{n_1(n_1+1)}{2} \qquad (4.4.)$$

d'autre part puisque $S_+ - S_- = n_0 n_1 - 2S_-$

on peut écrire : $$\tau_{bs} = 1 - \frac{2S_-}{n_0 n_1} \qquad (4.5.)$$

(Ce résultat intermédiaire est de quelque importance, puisque (4.5.) n'est autre que la forme particulière - correspondant au cas traité - de l'une des expressions données par Kendall (1948-1975) pour τ).
Portant (4.4.) dans (4.5.) on obtient :

$$\tau_{bs} = \frac{2}{n_0}\left[\frac{\Sigma y_{i1}}{n_1} - \frac{(n_0 + n_1 + 1)}{2}\right] \qquad (4.6.)$$

soit : $\tau_{bs} = r_b$

IV.2.2.3. Note sur le cas avec ex-æquo

Ne prenant pas en compte les ex-æquo de la même façon, τ_{bs} et r_b ont généralement des valeurs numériques différentes sauf dans des situations particulières dont certaines ont été étudiées par Cureton (1968). Aux cas envisagés par cet auteur on peut ajouter celui dans lequel il n'y a ni "bracket-tie", ni parmi les ex-æquo, de sujets qui occuperaient sur l'échelle Y les place n_0 et $n_0 + 1$ s'ils n'étaient pas ex-æquo.

Dans ce cas, le calcul de λ_i effectué plus haut reste valable si le $i^{ème}$ sujet classé 1 en X n'appartient pas à un ensemble d'ex-æquo ; dans le cas contraire on peut vérifier qu'on a :

$$\lambda_i' = n_0 - (\varepsilon_i - i)$$

où ε_i est la place qu'occuperait ce sujet s'il n'appartenait pas à un ensemble d'ex-æquo.

Alors :

$$S_- = \Sigma \lambda_i + \Sigma \lambda_i' = n_0 n_1 + \frac{n_1(n_1+1)}{2} - \left[\Sigma y_{i1} + \Sigma \varepsilon_i\right] \qquad (4.7.)$$

donc :

$$\tau_{bs} = \frac{2}{n_0}\left[\frac{\Sigma y_{i1} + \Sigma \varepsilon_i}{n_1} - \frac{(n_0+n_1+1)}{2}\right] \qquad (4.8.)$$

mais ici :

$$\overline{Y}_1 = \frac{\Sigma y_{i1} + \Sigma \varepsilon_i}{n_1}$$

d'où :

$$\tau_{bs} = r_b$$

On retrouverait, bien entendu, les résultats de IV.2.2. en raisonnant sur les sujets classés 0 en X.

V. CONCLUSION

C'est essentiellement en raison du type de données auxquelles il est applicable que nous dénommons τ_{bs} coefficient rang-bisérial.

Les possibilités d'utilisation semblent assez nombreuses : par exemple en expérimentation pharmacologique X peut être la présence ou l'absence d'un signe clinique et Y le nombre de doses, de jours de traitement... ; en psychothérapie X serait la présence ou l'absence de telle manifestation comportementale, Y le nombre d'entretiens avec le thérapeute. Plus généralement X peut être la variable indicateur d'un phénomène et Y la mesure ordinale du temps écoulé depuis le début des observations.

REMERCIEMENTS.

L'auteur remercie particulièrement le "referee" pour les suggestions qui ont permis d'écrire la version définitive de ce "papier".

BIBLIOGRAPHIE

[1] BROGDEN H.E., "A new coefficient : application to biserial correlation and to estimation of selective efficiency, *Psychometrika* 14 (1949), 169-182.

[2] CURETON E.E., "Rank biserial correlation", *Psychometrika* 21 (1956), 287-290.

[3] CURETON E.E., "Rank biserial correlation when ties are present", *Educ. and Psycho. Measurement*, 28 (1968), 77-79.

[4] DURBIN J., et STUART A., "Inversions and rank correlations coefficients", *J. Roy. Statis. Soc.*, 13 (1951), 303-309.

[5] GLASS G.V., "Note on rank biserial correlation", *Educ. and Psycho. Measurement*, 26 (1966), 623-631.

[6] KENDALL M.G., "A new measure of rank correlation", *Biometrika*, 30 (1938), 81 et suivantes.

[7] KENDALL M.G., "The treatment of ties in ranking problems", *Biometrika*, 33 (1945), 239 et suivantes.

[8] KENDALL M.G., "Rank correlation methods", Londres, Charles Griffin and Company ltd, 1ère édition 1948, 4ème édition 2ème tirage 1975.

[9] SIEGEL S., "Non parametric statistics for the behavarial sciences", New-York, Mac Graw and Hill, 1956.

[10] WHITFIELD J.W., "Rank correlation between two variables, one of which is ranked, the other dichotomous", *Biometrika*, 34 (1947), 292-296.