Tutorials in Quantitative Methods for Psychology
2009, Vol. 5(2), p. 51-58.

DOI: 10.20982/tqmp.05.2.p051

Le tau et le tau-b de Kendall pour la corrélation de variables ordinales simples ou catégorielles

L. Laurencelle,
Université du Québec à Trois-Rivières

Dans le langage de tous les jours, l'expression « corrélation entre deux variables » est entendue et bien comprise de manière générale : c'est un lien, un rapport de correspondance grâce auquel la variation d'un attribut peut être associée à la variation d'un autre attribut.

En fait, ce concept général de corrélation – on devrait plutôt employer le terme « association » – en recouvre trois qui, dans le langage plus rigoureux des interprétations scientifiques, ont chacun leur formulation mathématique et leur acception spécifiques. Ces trois types d'association sont :

une relation non spécifique : Y a-t-il un lien, une influence quelconque – causale, indirecte, mutuelle – entre deux caractéristiques observées ? Des exemples de ce type de relation abondent : le niveau moyen de réponse de deux ou plusieurs groupes, selon le traitement particulier auquel ont été soumis leurs participants, l'incidence de telle maladie selon la région géographique ou la race, la force musculaire selon le sexe, etc. Ces relations impliquent ordinairement une variable catégorielle (ou qualitative), tels le sexe, la condition expérimentale, la race, et une variable dépendante qui peut être ou mesurée (la force, le degré de réussite), ou observée (l'option politique, la présence de maladie). L'arsenal disponible pour l'évaluation statistique de ces relations est nombreux et diversifié, les outils les plus courants étant les tests *t* de Student, l'analyse de variance et le test du Khi-deux pour tableaux de fréquences. Laurencelle (2005) a considéré le cas d'une association non monotone entre deux variables continues.

une relation linéaire : La variation d'une grandeur *Y* donnée est-elle proportionnelle à la variation d'une autre grandeur *X*, peut-on établir une règle de correspondance linéaire entre les deux? Cette relation paradigmatique, représentée par l'équation « Y = bX + a », fournit souvent le premier pas vers une modélisation d'un phénomène et elle en constitue aussi l'aboutissement rêvé, parfois sous une forme augmentée telle que « $Y = f(X)$ », où *f* est une fonction, linéaire ou linéarisable, de *X*. Température et longueur d'une tige de fer, réussite scolaire et quotient intellectuel, fréquence cardiaque maximale et âge sont trois cas de relation linéaire simple à peu près démontrés. Le coefficient de corrélation linéaire, diversement attribué à Bravais, Galton et K. Pearson, en est la mesure directe, sous la forme simplifiée :

$$r \propto \Sigma\,(X_i - \mu_X)(Y_i - \mu_Y)\,.$$

La réécriture de l'équation « $Y = bX + a$ » sous la forme « $(Y - \mu_Y) = b(X - \mu_X)$ » montre à l'évidence que cette corrélation quantifie le degré de correspondance proportionnelle entre *X* et *Y* et qu'elle est donc intimement solidaire de la métrique de ces deux variables. L'analyse de régression, la régression polynomiale et le modèle linéaire général comptent parmi les techniques d'évaluation privilégiées pour ce type de relation.

une relation monotone : Intermédiaire entre la relation linéaire décrite ci-dessus et une relation non spécifique, qui n'implique aucune concordance régulière entre *X* et *Y*, s'intercale la relation dite monotone. Y a-t-il une fatigue, une diminution d'énergie chez les travailleurs du premier jusqu'au cinquième jour d'une semaine de travail? Les enfants dont la diète comporte une dose plus importante de fruits et légumes sont-ils moins sujets à certaines maladies? L'incidence d'accidents vasculocérébraux augmente-t-elle avec le cholestérol sanguin total ? Les relations de ce type se subdivisent en deux sous-catégories, selon la nature des variables concernées :

- soit les variables *X* et *Y* concernées sont toutes deux mesurées et reflètent des grandeurs (continues), alors que la relation supposée entre elles est *monotone* plutôt que métrique (p. ex. l'espérance de vie *vs* le revenu moyen);
- soit au moins l'une des variables est ordinale (p. ex. l'échelle de Mohs pour la dureté d'un matériau) ou catégorielle ordonnée, la variable catégorielle regroupant

ordinairement plus d'un élément (p. ex. le niveau de scolarité).

Par relation monotone, on entend une correspondance entre les variables X et Y telle qu'une inégalité dans une variable entraîne une inégalité dans l'autre, la relation étant positive si $X_i > X_j$ entraîne habituellement $Y_i > Y_j$ ou négative si $X_i > X_j$ entraîne $Y_i < Y_j$[1]. Ce type de relation, sans modèle spécifié et qui concerne des variables sans métrique ou à métrique floue (pour ainsi dire), n'a guère été achalandé par les chercheurs, et les techniques disponibles sont peu nombreuses et comparativement peu connues. D'un côté, suite à Jonckeere (1954), Barlow, Bartholomew, Bremmer et Brunk (1972) mirent au point un test de « variation monotone », applicable dans le contexte de l'analyse de variance, l'une des variables étant catégorisée (voir Laurencelle 1993, Laurencelle et Dupuis 2000, Mukerjee 1988). De l'autre, Kendall (Kendall et Gibbons 1990) a consacré son fameux indice « tau », lequel quantifie le degré de concordance monotone entre deux variables ordinales, qu'elles soient continues (indice τ, ou tau) ou catégorielles (indice τ_b, ou tau-b).

Comme le fera voir l'exposé plus loin, le tau de Kendall a, sur la corrélation r de Pearson, deux avantages décisifs, celui de transcender la métrique des variables mesurées et celui d'écarter la nécessité d'un modèle paramétrique de relation (comme le modèle linéaire), ce qui en fait un indice authentiquement non paramétrique, approprié aux variables de type ordinal. Tel n'est pas le cas du coefficient « rho » de Spearman (noté ici r_S), lequel invoque un modèle linéaire strict, c'est-à-dire un calcul ordinaire du r de Pearson appliqué sur les rangs des variables considérées.

Quant à la corrélation entre variables ordinales catégorielles, elle mérite d'être ré-examinée avec attention. Dans un exemple type (tel que celui présenté plus loin, voir Calcul du τ_b), des observations X,Y en nombre important (n) sont réparties en quelques catégories d'une variable (k_X) et de l'autre (k_Y), le nombre moyen de valeurs partagées ou égales dans chaque variable (n/k_X, n/k_Y) étant élevé. Ce contexte statistique conduit naturellement à la constitution d'un tableau croisé, d'ordre $k_X \times k_Y$, qui synthétise l'information : chaque cellule contient alors le nombre d'observations inscrites concomitamment dans une catégorie d'une variable (X_i) et de l'autre (Y_j). Or, l'outil traditionnel appliqué au traitement d'un tableau croisé de fréquences tel que celui-là est le test du khi-deux, le khi-deux d'interaction, qui permet de rejeter l'hypothèse de l'indépendance stochastique des deux variables mesurées. Toutefois, comme l'ont bien montré Burr (1960) et Somers (1962), grâce à des exemples et contre-exemples, le khi-deux d'interaction ne convient pas à cet usage : la sanction que donne le khi-deux est générale et ne porte pas spécifiquement sur l'hypothèse d'une relation monotone entre les variables X et Y. Pour cette raison, un khi-deux significatif n'indique pas forcément la présence d'une relation monotone, et la présence d'une relation monotone statistiquement sérieuse (p. ex. d'après un argument de probabilité direct) pourrait passer inaperçue au khi-deux en raison de sa faible puissance. Le test approprié, le tau-b (Kendall et Gibbons 1990), est une variante simple du tau, qui tient compte du caractère catégoriel des variables observées et pour lequel un calcul à partir du tableau croisé de fréquences est aussi suggéré. Le lecteur trouvera une variante asymétrique du tau-b dans Somers (1962).

Les variables catégorielles, qu'on désigne parfois « qualitatives », apparaissent principalement dans les données d'enquête et de sondage : épidémiologie, enquête de consommation, opinion publique, et leur traitement coutumier passe par le progiciel SPSS ou par d'autres moyens semblables. Les interdépendances possibles entre variables sont ordinairement testées par le khi-deux d'interaction, *même dans les cas où les deux variables concernées sont ordonnées* : une pléthore de publications peut être citée à l'appui. Or, tel que mentionné précédemment et dans ce contexte :

1- l'obtention d'un khi-deux significatif indique que les deux variables sont statistiquement non indépendantes[2], non pas qu'elles ont un rapport de relation monotone l'une avec l'autre ;

2- l'obtention d'un khi-deux non significatif n'implique pas qu'il n'y ait pas de relation monotone significative entre les deux variables.

C'est pourquoi il nous est apparu important de présenter à nouveau cet outil particulier qu'est le tau de Kendall, et plus spécialement de réactualiser le tau-b pour la corrélation entre deux variables ordinales catégorielles.

Calcul du τ, avec exemple

La formule classique du τ de Kendall, applicable à des variables ordinales simples, sans catégorisations ni valeurs égales, est :

$$\tau = (C - D) / \binom{n}{2}, \qquad (1)$$

où C = nombre d'inégalités concordantes [p.ex. $X_i > X_j$ et $Y_i > Y_j$] ; D = nombre d'inégalités discordantes [p.ex. $X_i > X_j$ et $Y_i < Y_j$] ; n = nombre de paires X,Y.

[1] Le verbe « entraîner » est proposé ici dans une acception descriptive, statistique, et ne présuppose pas le concept de causalité ou d'influence.

[2] Dans le test du khi-deux d'interaction, la composante additive du test, « $(n_{i,j} - n_i \times n_j)^2 / (n_i \times n_j)$ », reste inchangée sous toute permutation des rangées ou des colonnes du tableau, de sorte que la valeur du khi-deux reste la même, la corrélation (monotone) dépendant quant à elle de l'ordre des catégories tel qu'observé.

La différence (C – D), au numérateur de (1), peut être obtenue par la somme $\Sigma_{i,j}$ signe$\{(X_i - X_j)(Y_i - Y_j)\}$, effectuée sur les $\binom{n}{2} = \frac{1}{2}\, n(n-1)$ paires possibles d'observations, où signe(z) = –1, 0 ou 1 selon que $z < 0$, $z = 0$ ou $z > 0$ respectivement.

Une autre méthode, indiquée par exemple dans Siegel et Castellan (1988) pour le calcul manuel, consiste en deux étapes. En premier, il s'agit de placer la série *X*,*Y* en ordre croissant des valeurs *X*. En second, pour chacune des observations Y_i, de $i = 1$ à $i = n-1$, il faut compter le nombre de valeurs Y_j suivantes qui sont plus grandes (c.-à-d. $Y_j > Y_i$ pour $j > i$), moins le nombre de valeurs plus petites ($Y_j < Y_i$), le total donnant C – D. La même formule (1) s'applique. L'exemple 1 illustre les calculs.

Exemple 1. Données ordinales distinctes

X	10	4	16	5	13	14
Y	17	14	20	8	11	23

L'exemple nous met en présence de $n = 6$ paires *X*,*Y* ; les données de chaque variable sont toutes distinctes. Ces $n = 6$ valeurs donnent lieu à $\binom{n}{2} = 15$ comparaisons. La liste des comparaisons suit :

compar. en X	*compar. en* Y	*concord.*	*discord.*
10 > 4	17 > 14	T	
10 < 16	17 < 20	T	
10 > 5	17 > 8	T	
10 < 13	17 > 11		T
10 < 14	17 < 23	T	
4 < 16	14 < 20	T	
4 < 5	14 > 8		T
4 < 13	14 > 11		T
4 < 14	14 < 23	T	
16 > 5	10 > 8	T	
16 > 13	20 > 11	T	
16 > 14	20 < 23		T
5 < 13	8 < 11	T	
5 < 14	8 < 23	T	
13 < 14	11 < 23	T	
		C =11	D = 4

L'analyse donne C = 11 comparaisons concordantes, c'est-à-dire des comparaisons pour lesquelles l'inégalité observée en *X* est de même signe que celle observée en *Y*, et D = 4 comparaisons discordantes ; notons que $C + D = \binom{n}{2} = 15$. Appliquant la formule (1), nous obtenons donc :

$$\tau = (11 - 4) / \binom{6}{2}$$
$$= 7 / 15$$
$$= 0{,}467 .$$

Pour le procédé « manuel » suggéré par Siegel et Castellan (1988), nous devons d'abord placer les observations *X* en ordre de valeurs croissantes, accompagnées des Y_i correspondantes. Ensuite, pour chaque Y_i, il nous faut ensuite compter, en plus ou en moins, le nombre d'observations Y_j suivantes (pour lesquelles $j > i$) selon qu'elles sont plus fortes ou moins fortes que Y_i : le total donne C – D, tel qu'illustré par le tableau suivant[3].

X	Y	Σ *signe*$(Y_j - Y_i)$
4	14	3 – 2
5	8	4 – 0
10	17	2 – 1
13	11	2 – 0
14	23	0 – 1
16	20	
		C–D = 11– 4 = 7

Prenons, par exemple, la première valeur de *Y*, $Y_1 = 14$; dans la série des Y_j qui la suivent, les valeurs 17, 23 et 20 sont supérieures (> 14), tandis que 8 et 11 sont inférieures, d'où le compte net est de 1, La compilation pour Y_2 et les *Y* suivantes donne un total (C – D) de 7. L'application de la formule (1) fournit encore :

$$\tau = 7 / 15$$
$$= 0{,}467 .$$

Nous considérons plus loin la distribution de probabilité de cet indice et la procédure de test qui lui convient.

Calcul du τ_b, avec exemples

L'indice tau-b (noté τ_b) est une extension du τ de Kendall pour données ordinales non distinctes, c'est dire qu'il s'applique de façon particulièrement idoine aux variables ordinales catégorielles. Pour le calcul, la règle des signes, $\Sigma_{i,j}$ signe$\{(X_i - X_j)(Y_i - Y_j)\}$, s'applique encore, le cas de valeurs égales, soit $X_i = X_j$ ou $Y_i = Y_j$, aboutissant à un résultat de 0.

La formule à retenir est :

$$\tau_b = (C - D) / \text{Den} \qquad (2)$$

où :

$$\text{Den} = \tfrac{1}{2}\sqrt{\left[n(n-1) - U_X\right]\left[n(n-1) - U_Y\right]} \qquad (3)$$

et :

$$U_X = \Sigma\, u_X(u_X - 1) , \qquad (4)$$
$$U_Y = \Sigma\, u_Y(u_Y - 1) ;$$

les symboles « u_X » et « u_Y » dénotent le nombre d'apparitions de chaque valeur *X* et chaque valeur *Y* respectivement. Noter que la formule (1) est un cas particulier de celle-ci, où $u_X = u_Y = 1$ pour tous les *X* et tous les *Y* (chaque valeur n'apparaissant qu'une fois). Nous

[3] L'astuce ici vient de ce que les données ont été ordonnées selon *X*, de sorte qu'il n'est pas utile de tenir compte du signe des inégalités en *X* (qui sont alors toutes positives).

présentons deux exemples de calcul, un premier, plus simple, à partir d'une courte série *X*,*Y*, et un second, basé sur un tableau croisé de fréquences.

Calcul sur une série (exemple 2). L'exemple 2, ci-dessous, servira d'illustration.

Exemple 2. Données ordinales avec égalités

X	10	18	18	20	26	36
Y	12	19	12	31	29	33

Le procédé de Siegel et Castellan (1988), hélas, ne s'applique pas ici, en raison des permutations possibles des paires comportant des valeurs égales. Il faut plutôt recourir laborieusement aux comparaisons par paires. Parmi les $\binom{6}{2}$ = 15 paires générées, la série *X*,*Y* de l'exemple 2 contient C = 12 paires concordantes, D = 1 paire discordante et 2 paires avec égalités, d'où résulte le numérateur C – D = 11. Pour le dénominateur (3), les quantités U_X et U_Y sont nourries par le nombre de valeurs (*X* ou *Y*) apparaissant plus d'une fois. Ici, en *X*, seule la valeur « 18 » apparaît 2 fois, i.e. $u_{18} = 2$, d'où $U_X = 2\times(2–1) = 2$; de même pour la valeur « 12 » en *Y*, d'où $U_Y = 2$. Les autres valeurs de *X* ou *Y*, n'apparaissant qu'une fois (i.e. pour *X* = 10, $u_{10} = 1$), n'ajoutent rien à leur somme respective U_X et U_Y. Le dénominateur devient donc :

$$\text{Den} = \tfrac{1}{2}\sqrt{(6\times5-2)(6\times5-2)} = 14.$$

Nous obtenons enfin :

$$\tau_b = 11 / 14$$

Exemple 3. Données ordinales catégorielles

		Réseau social				
		Aucun	Proches	Proches et amis occasionnels	Groupe régulier	Total
Activité physique	Jamais	*5*	*3*	3	*2*	13
	1-2 fois/ an	*8*	*7*	4	*1*	20
	1-2 fois/ mois	*3*	*8*	8	*2*	21
	1 fois / sem	5	5	*6*	4	20
	2 jrs ou plus / sem	2	0	8	2	12
	Chaque jour	5	0	5	4	14
	Total	28	23	35	15	*n* = 100

= 0,786 .

Calcul sur un tableau X × Y (exemple 3). Le calcul du τ_b à partir d'un tableau croisé de fréquences met en oeuvre les mêmes principes que ceux pratiqués dans l'exemple précédent, seul le format de la compilation change. Nous baserons notre explication sur l'exemple 3, ci-dessous. Au total, 100 (= *n*) personnes ont été interrogées quant à leur réseau social (au travail et hors travail) et leur pratique d'activités physiques. En admettant que les catégories d'activité physique croissent en intensité (depuis « jamais » jusqu'à « chaque jour »), tout comme les catégories du réseau social en importance (depuis « aucun (réseau) » jusqu'à « groupe régulier »), nous nous demandons s'il y a un lien de corrélation entre l'une et l'autre variable.

Le tableau, comme on voit, se présente avec des catégories de « valeurs », croissantes de gauche à droite et de haut en bas : le dénombrement des inégalités concordantes et discordantes en tire profit. Prenons, par exemple, le cas des personnes pratiquant « 1 fois / sem » et ayant comme réseau des « Proches et amis occasionnels » : leur nombre est de 6. Alors, dans le tableau, toutes les observations qui se situent en haut et à gauche sont concordantes (leur nombre est 5 + 3 + 8 + 7 + 3 + 8 = 34), et celles situées en haut et à droite sont discordantes (leur nombre est 2 + 1 + 2 = 5) : la portion nette d'inégalités relative à cette cellule est donc 6 × (34 – 5) = 174.

Le calcul systématique peut être noté comme suit (depuis la 2e ligne du tableau) :

8 × (0–8) +	7 × (5–5) +	4 × (8–2) +	1 × (11–0) =	–29
3 × (0-20) +	8 × (13-10) +	8 × (23–3) +	2 × (30–0) =	184
5 × (0–38) +	5 × (16–20) +	*6 × (34–5)* +	4 × (49–0) =	160
2 × (0–53) +	0 × (21–30) +	8 × (44–9) +	2 × (65–0) =	304
5 × (0–63) +	0 × (23–40) +	5 × (46–11) +	4 × (75–0) =	160
			C–D(total)=	779

Le numérateur du τ_b est donc C – D = 779. Quant au dénominateur, il faut d'abord établir les quantités U_X et U_Y, ce à partir des totaux de rangées (*X*) et colonnes (*Y*) du tableau croisé ; ces totaux indiquent le nombre de « valeurs égales » pour chaque valeur *X* et *Y*. Nous avons donc :

$$U_X = 13\times(13–1) + 20\times(20–1) + \ldots + 14\times(14–1) = 1650,$$
$$U_Y = 28\times(28–1) + 23\times(23–1) + \ldots + 15\times(15–1) = 2594.$$

d'où le dénominateur (3) s'obtient par :

$$\text{Den} = \tfrac{1}{2}\sqrt{(100\times99-1650)(100\times99-2594)}$$

$$\approx 3881{,}833 .$$

La valeur du tau-b est enfin :

$$\tau_b = 779 / 3881{,}833$$
$$\approx 0{,}2007 .$$

Nous considérons maintenant les propriétés statistiques

(distribution et moments) des indices τ et τ_b, et présentons des tests de significativité appropriés.

Distribution, moments, test et approximation pour τ

Le numérateur de τ, s = C – D, peut prendre ½n(n–1) + 1 valeurs différentes, qui courent de s = –½n(n–1) à s = ½n(n–1) par incréments de 2 ; par exemple, pour n = 3, s peut prendre les valeurs –3, –1, 1, 3. De plus, le nombre total d'arrangements des séries X,Y générant ces valeurs de s est n!, chacun ayant une probabilité de 1 / n!. Kendall et Gibbons (1990) détaillent le procédé récursif suivant, qui permet de générer la distribution de probabilité de τ sous des permutations au hasard des séries. Soit n = 2, les valeurs s = –1 et 1 (correspondant aux séries 2-1 et 1-2), et leurs fréquences relatives { 1 1 }. Ainsi, la valeur s = –1 (correspondant à τ = –1) a une probabilité de 1 / 2! = 0,50. Pour générer les fréquences pour des séries de n à partir de n–1, il faut superposer n fois la série antécédente, en la décalant chaque fois d'une position à droite, puis additionner les fréquences qui sont en vis-à-vis. Ainsi, pour n = 3 (depuis n – 1 = 2), nous avons :

1	1		
	1	1	
		1	1
1	2	2	1

Les fréquences « 1 2 2 1 », qui représentent 3! = 6 arrangements, correspondent aux valeurs s = –3, –1, 1, 3. Pour aller à n = 4, nous reprenons pareillement la série de n–1 = 3, et :

1	2	2	1			
	1	2	2	1		
		1	2	2	1	
			1	2	2	1
1	3	5	6	5	3	1

Ces fréquences, de total 4! = 24, correspondant à s = –6, –4, –2, 0, 2, 4, 6. Kendall et Gibbons (*op. cit.*) proposent un second procédé récursif de calcul, que nous omettons ici.

Les moments de la distribution de τ sont :

$$\mu = 0 \tag{5a}$$

$$\sigma^2 = (4n + 10) / 9n(n–1) \tag{5b}$$

$$\gamma_1 = 0 \tag{5c}$$

$$\gamma_2 = -\left(\frac{6}{5}\right)^2 \frac{6n^3 + 21n^2 + 31n + 31}{n(n-1)(2n+5)^2} \tag{5d}$$

La distribution est évidemment symétrique (γ_1 = 0). Son indice d'aplatissement, γ_2, avoisine de –2/(n–1), se rapprochant rapidement de 0 avec n croissant. Ces deux caractéristiques promettent une approximation normale d'un certain succès.

L'approximation de la distribution de τ par une normale standard passe par la formule :

$$z = [\ \tau \pm K/n(n–1)] / \sigma\ ,\ \ K = 2, \tag{6}$$

laquelle comporte une correction de continuité (voir aussi Burr 1960). Best (1973), qui a étudié cette approximation, propose qu'elle est déficiente pour $n \le 40$ et il publie des valeurs critiques exactes allant jusqu'à n = 100. Nous avons étudié une approximation plus serrée, grâce à la loi Bêta symétrique, $\beta(p,p)$ qui reproduit à peu près l'aplatissement de τ avec p = 1,39892n – 1,56354 ; à notre surprise, l'approximation par la Bêta s'est révélée décevante, voire parfois inférieure à la normale. Nous nous accordons partiellement avec Best, à la nuance suivante. Les valeurs critiques exactes semblent requises (aux seuils de signification habituels) pour $n \le 29$. Pour $n \ge 30$, l'approximation (6) s'appliquerait mais avec une correction utilisant K = 3 plutôt que K = 2.[4] Dès $n \ge 100$, l'usager peut utiliser sans crainte l'approximation (6) telle que présentée. À toutes fins utiles, nous incluons en annexe notre propre table de valeurs critiques exactes du τ, pour n = 4 à 100 et les seuils de signification 0,10, 0,05, 0,025, 0,01 et 0,005.

L'exemple 1, ci-dessus, comportait une série de n = 6 données X,Y, avec un indice τ = 0,467 (= 7 / 15). Pour n = 6, les moments du τ sont μ = 0, σ^2 = 34 / 270 ≈ 0,1259 ($\sigma \approx 0{,}355$), γ_1 = 0 et $\gamma_2 \approx -0{,}3769$. Comme les fréquences ponctuelle et cumulative pour s = 7 sont 49 et 671, le τ = 7 / 15 ≈ 0,467 obtient une probabilité exacte de (720 – 671 + 49) / 6! ≈ 0,136. La table 1, en annexe, propose ±0,867 comme valeurs critiques au seuil de 5 % bilatéral. Quant à l'approximation normale (équation 6, avec K = 3), guère utile ici en raison du petit n = 6, elle donne :

$$z = (\ 0{,}467 – 3/(6\times5)\ ) / 0{,}355$$
$$\approx 1{,}034,$$

valeur dont la probabilité extrême (sous la loi normale standard) est 0,151.

Moments, test et approximation pour τ_b

Considérablement plus complexe, la distribution du tau-b est aussi moins documentée. Notant encore s = C – D pour le numérateur de l'indice τ_b (2), Burr (1960) détaille la distribution de probabilité pour tous les arrangements (avec égalités) possibles pour $n \le 6$, tandis que Sillito (1947) le fait pour $n \le 10$ mais en ne considérant que des égalités de 2 ou 3 composantes des séries. Valz, McLeod et Thompson (1995) tentent, quant à eux, d'établir une fonction génératrice approximative de la distribution du τ_b, sans grand succès, l'approximation normale restant la première en lice pour des tailles n suffisantes.

Rappelant la formule (2) sous la forme τ_b = s / Den, Kendall et Gibbons (1990) démontrent les deux premiers

[4] La platykurtose (γ_2 < 0) encore présente dans la distribution explique sans doute en partie ce changement.

moments :

$$\mu(s) = 0 \quad (7a)$$

$$\text{var}(s) = [n(n–1)(2n+5) – V_X – V_Y]/18 + W_X W_Y / [9n(n–1)(n–2)] + U_X U_Y /[2n(n–1)] \quad (7b)$$

où :

$$U_X = \Sigma\, u_X(u_X–1) \; ; \; U_y = \Sigma\, u_Y(u_Y–1) \quad (8a)$$

$$V_X = \Sigma\, u_X(u_X–1)(2u_X + 5) \; ; \; V_Y = \Sigma\, u_Y(u_Y–1)(2u_Y + 5) \quad (8b)$$

$$W_X = \Sigma\, u_X(u_X–1)(u_X–2) \; ; \; W_Y = \Sigma\, u_Y(u_Y–1)(u_Y–2) \quad (8c)$$

d'où, évidemment :

$$\mu (\tau_b) = 0 \quad (9a)$$

$$\text{var} (\tau_b) = \text{var}(s) / (\text{Den})^2 \quad (9b)$$

Comme on peut imaginer et hormis les exceptions déjà mentionnées, il n'existe pas et il n'est pas raisonnable de demander de tables de valeurs critiques exactes pour des séries de tailles n utiles[5]. Le critère distributionnel (par permutations aléatoires de la série Y, voir p. ex. Edgington 1980 ou Laurencelle 2001) peut suppléer aux tables dans le cas de tailles n modestes ; au delà, l'approximation normale (6), avec $K = 2$, est recommandée.

Pour l'exemple 2, avec une série X,Y de $n = 6$ éléments, les ingrédients (8a)-(8c) de la variance sont : $U_X = 2$, $U_Y = 2$, $V_X = 18$, $V_Y = 18$, $W_X = 0$, $W_Y = 0$, d'où var(s) = 26,4. Utilisant la valeur déjà calculée du dénominateur, Den = 14,0, nous obtenons var(τ_b) ≈ 26,4 / $14{,}0^2$ ≈ 0,1347 et $\sigma \approx 0{,}367$. La valeur observée du tau-b était 0,786. L'évaluation de la probabilité extrême par voie Monte Carlo (basée sur 50 000 permutations aléatoires) indique une probabilité extrême p = 0,022. Pour l'approximation normale (non recommandée pour $n < 30$), nous calculons z = (0,786 – 2/30) / 0,367 ≈ 1,960 et $p(z) \approx 0{,}025$.

Le traitement de l'exemple 3, sur tableau croisé de fréquences, entraîne de plus lourds calculs. Ayant déjà établi plus haut U_X = 1650, U_Y = 2594 et Den ≈ 3881,833, nous totalisons V_X = 68610, V_Y = 161178, W_X = 26880 et W_Y = 68916, d'où var (s) ≈ 100412,32 et var (τ) ≈ 0,006664, d'où $\sigma \approx 0{,}0816$. L'approximation normale fournit z = (0,2007 – 2 / 9900) / 0,0816 ≈ 2,457, d'où p (z) ≈ 0,007 ; sur 50000 permutations au hasard effectuées sur la série des Y, 350 obtiennent une valeur $\tau \geq 0{,}2007$, d'où p = 0,007 aussi. Notons ici que le khi-deux d'interaction, établi sur le tableau de l'exemple 3, vaut $\chi^2 = 22{,}62$, avec 15 degrés de liberté, la probabilité extrême étant $p \approx 0{,}093$. Nous avons donc ici le cas d'une relation monotone significative (avec $p < 0{,}01$) entre X et Y, relation que le khi-deux ne débusque pas.

[5] Des valeurs critiques exactes devraient tenir compte de tous les assortiments possibles (en termes de valeurs égales ou inégales) des n valeurs observées en X et en Y, un défi titanesque. Par exemple, pour n = 4, on compte 4 partitions non triviales (soit 31, 22, 211, 1111), où par exemple « 31 » indique 3 valeurs égales + 1 valeur autre , et 7 « compositions » (soit 31, 13, 22, 211, 121, 112, 1111), où « 31 » indique que les 3 valeurs égales sont les plus faibles, au contraire de « 13 » : d'où la nécessité de préparer 7^2 = 49 séries critiques dans ce cas. Pour n = 5 et 6, nous obtenons respectivement 6 et 10 partitions, d'où 225 et 784 séries. Enfin, pour n = 10 par exemple, il y a 41 partitions, 457 compositions et 208849 séries requises!

Discussion et conclusion

Les variables de niveau linéaire, pour lesquelles les différences de valeurs sont mesurables (p. ex. $X_1 – X_2 > X_3 – X_4$), ont aussi une capacité ordinale (p. ex. $X_1 > X_2$), de sorte que les techniques de corrélation de Pearson (r), Spearman (r_S) *et* Kendall (τ) s'y appliquent toutes. Kendall et Gibbons (1990 ; voir aussi Kendall et Stuart, 1979) donnent les équivalences approximatives entre ces trois statistiques :

$$\tau \approx \tfrac{1}{2}\pi \sin^{-1} r \,, \quad (10a)$$

$$r \approx 2\sin(\pi r_S/6) \,, \quad (10b)$$

d'où d'autres équivalences peuvent être dérivées. Kendall et Gibbons présentent aussi une variante de τ dans le cas où la population X,Y serait d'ordre linéaire et binormale, ce qui donne lieu à un test normal approximatif plus puissant que (6). Il reste que, pour des variables X,Y proprement ordinales, par conséquent non normales, seul le tau de Kendall s'applique légitimement, et l'utilisation du r ou du r_S ne peut y être considérée qu'abusive.

Agresti (1976) passe en revue divers indices de corrélation appliqués au traitement de données ordinales catégorielles, incluant le tau-b, le r_S de Spearman et un indice élémentaire « gamma », γ = (C – D)/(C + D), la notation étant celle de notre formule (1). Son étude lui permet de conclure à la supériorité du tau-b, lequel serait moins affecté par des changements dans la catégorisation des variables et conserverait davantage sa convergence vers sa valeur paramétrique.

Le traitement des variables ordinales catégorielles, particulièrement pour la corrélation entre deux variables disposées en tableau, est traditionnellement problématique. Le hic tient à l'usage du khi-deux d'interaction sur ces données, tel que dénoncé par Burr (1960) et Somers (1962). Le khi-deux est un test « omnibus » applicable sur un tableau croisé de fréquences, test qui rejette l'hypothèse nulle (indépendance stochastique des variables X et Y) en faveur de toutes formes de dépendance, incluant la dépendance monotone. Pour illustration, prenons le tableau simpliste suivant (chaque ligne est identifiée par une lettre) :

a	*2*	*0*	*0*	*0*
b	0	2	0	0
c	0	0	2	0
d	0	0	0	2

Tel que présenté et en supposant que les rangées et colonnes sont en ordre (croissant) des variables, le tableau de fréquences 4 × 4 génère une valeur $\tau_b = 1$. Les rangées dénotées « a », « b », « c » et « d » peuvent aussi être

permutées, ce de 4! = 24 façons, ce qui donne lieu aux valeurs τ_b suivantes, chacune associée à sa probabilité extrême :

Comme on voit, les 24 réarrangements de ce petit tableau produisent jusqu'à 7 valeurs distinctes de corrélation ordinale[6], *tandis que tous aboutissent à un seul et même khi-deux d'interaction*, soit $\chi^2 = 24{,}00$, doté de 9 degrés de liberté et de

τ_b =	*1*	*2/3*	*1/3*	*0*	*–1/3*	*–2/3*	*–1*
$p(\tau_b) \approx$	0,0004	0,019	0,161	0,533	0,161	0,019	0,0004
liste de permutations	abcd	abdc acbd bacd	acdb adbc badc bcad cabd	adcb bcda bdac cadb cbad dabc	bdca cbda cdab dacb dbac	cdba dbca dcab	dcba

probabilité $p \approx 0{,}004$. Conclure, sur la foi d'un test khi-deux, qu'il y a ou qu'il n'y a pas de corrélation dans un tableau croisé de fréquences serait donc commettre une grande imprudence.

Pour l'étude de corrélation de variables ordinales, catégorisées ou non, la technique de Kendall apparaît incontournable, et le traitement par le khi-deux ne peut être vu que comme un pis-aller dangereux.

Références

Agresti, A. (1976). The effect of category choice on some ordinal measures of association. *Journal of the American Statistical Association*, 71, 49-55.

Barlow, R.E., Bartholomew, D.J., Bremmer, J.M., Brunk, H.D. (1972). *Statistical inference under order restrictions*. New York : Wiley.

Best, D.J. (1973). Extended tables for Kendall's tau. *Biometrika*, 60, 429-430.

Burr, E.J. (1960). The distribution of Kendall's score S for a pair of tied rankings. *Biometrika*, 47, 151-171.

Edgington, E.S. (1980). *Randomization tests*. New York : Marcel Dekker.

Jonckeere, A.R. (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika*, 41, 133-145.

Kendall, M., et Gibbons, J.D. (1990). *Rank correlation methods (5e édition)*. New York : Oxford University Press.

Kendall, M, Stuart, A. (1979). *The advanced theory of statistics. Volume 2. Inference and relationship (4e édition)*. New York : Macmillan.

Laurencelle, L. (1993). Deux tests de variation monotone pour l'analyse de variance. *Lettres Statistiques*, 9, 69-78.

Laurencelle, L., Dupuis, F.A. (2000). *Tables statistiques expliquées et appliquées (2e édition)*. Québec : Le Griffon d'argile.

Laurencelle, L. (2001). *Hasard, nombres aléatoires et méthode Monte Carlo*. Québec : Presses de l'Université du Québec.

Laurencelle, L. (2005). Un indice d'association entre deux variables continues. *Lettres Statistiques*, 12, 81-98.

Mukerjee, H. (1988). Order restricted inference in a repeated measure model. *Biometrika*, 75, 616-617.

Sillito, G.P. (1947). The distribution of Kendall's τ coefficient of rank correlation in rankings containing ties. *Biometrika*, 34, 36-40.

Siegel, S., Castellan, N.J. (1988). *Nonparamtric statistics for the behavioral sciences (2e édition)*. New York : McGraw-Hill.

Somers, R.H. (1962). A new asymmetric measure of association for ordinal variables. *American sociological review*, 27, 799-811.

Valz, P.D., McLeod, A.I., Thompson, M.E. (1995). Cumulant generating function and tail probability approximations for Kendall's score with tied rankings. *Annals of statistics*, 23, 144-160.

Article received May 5th, 2009
Article accepted September 19th, 2009

Appendices follows.

[6] Par comparaison, les permutations des 8 données individuelles, au nombre de 8! / (2!)⁴ = 2520, donnent lieu à 31 valeurs distinctes de τ_b et à 21 valeurs du r_S de Spearman.

Annexe. Valeurs critiques du τ de Kendall (n valeurs distinctes en X et en Y)

n	0,10	0,05	0,025	0,01	0,005
3	-	-	- -	-	
4	1	1	- -	-	
5	0.800	0.800	1 1	-	
6	0.600	0.733	0.867	0.867	1
7	0.524	0.619	0.714	0.810	0.905
8	0.429	0.571	0.643	0.714	0.786
9	0.389	0.500	0.556	0.667	0.722
10	0.378	0.467	0.511	0.600	0.644
11	0.345	0.418	0.491	0.563	0.600
12	0.303	0.394	0.455	0.545	0.576
13	0.308	0.359	0.436	0.513	0.563
14	0.275	0.363	0.407	0.473	0.516
15	0.276	0.333	0.390	0.467	0.505
16	0.250	0.317	0.383	0.433	0.483
17	0.250	0.309	0.368	0.426	0.471
18	0.242	0.294	0.346	0.412	0.451
19	0.228	0.287	0.333	0.392	0.439
20	0.221	0.274	0.326	0.379	0.421
21	0.210	0.267	0.314	0.371	0.410
22	0.203	0.264	0.307	0.359	0.394
23	0.202	0.257	0.296	0.352	0.391
24	0.196	0.246	0.290	0.341	0.377
25	0.193	0.240	0.287	0.333	0.367
26	0.188	0.237	0.280	0.329	0.360
27	0.179	0.231	0.271	0.322	0.356
28	0.180	0.228	0.265	0.312	0.344
29	0.172	0.222	0.261	0.310	0.340
30	0.172	0.218	0.255	0.301	0.333
31	0.166	0.213	0.252	0.295	0.325
32	0.165	0.210	0.246	0.290	0.323
33	0.163	0.205	0.242	0.288	0.314
34	0.159	0.201	0.237	0.280	0.312
35	0.156	0.197	0.234	0.277	0.304
36	0.152	0.194	0.232	0.273	0.302
37	0.150	0.192	0.228	0.267	0.297
38	0.149	0.189	0.223	0.263	0.292
39	0.147	0.188	0.220	0.260	0.287
40	0.144	0.185	0.218	0.256	0.285
41	0.141	0.180	0.215	0.254	0.280
42	0.141	0.178	0.213	0.250	0.275
43	0.138	0.176	0.209	0.247	0.274
44	0.137	0.173	0.207	0.243	0.268
45	0.135	0.172	0.204	0.240	0.267
46	0.132	0.169	0.202	0.239	0.264
47	0.132	0.167	0.199	0.236	0.260
48	0.129	0.167	0.197	0.232	0.257
49	0.129	0.163	0.196	0.230	0.253
50	0.127	0.162	0.192	0.228	0.25
51	0.126	0.161	0.191	0.225	0.249

n	0,10	0,05	0,025	0,01	0,005
52	0.124	0.158	0.189	0.223	0.246
53	0.123	0.157	0.187	0.221	0.244
54	0.122	0.156	0.185	0.219	0.241
55	0.121	0.154	0.182	0.216	0.239
56	0.119	0.152	0.181	0.214	0.236
57	0.118	0.152	0.179	0.212	0.234
58	0.117	0.149	0.177	0.210	0.232
59	0.116	0.148	0.176	0.209	0.230
60	0.115	0.147	0.174	0.207	0.228
61	0.114	0.145	0.173	0.204	0.226
62	0.113	0.144	0.172	0.203	0.224
63	0.112	0.143	0.171	0.201	0.223
64	0.111	0.142	0.169	0.199	0.220
65	0.110	0.140	0.167	0.198	0.218
66	0.110	0.139	0.166	0.196	0.217
67	0.108	0.139	0.164	0.195	0.215
68	0.107	0.137	0.163	0.193	0.213
69	0.107	0.136	0.162	0.192	0.211
70	0.106	0.135	0.161	0.190	0.210
71	0.105	0.134	0.160	0.189	0.209
72	0.104	0.133	0.158	0.188	0.207
73	0.104	0.132	0.158	0.186	0.205
74	0.103	0.131	0.156	0.185	0.204
75	0.102	0.130	0.155	0.183	0.203
76	0.101	0.129	0.154	0.182	0.201
77	0.100	0.129	0.153	0.181	0.200
78	0.100	0.128	0.152	0.179	0.199
79	0.099	0.127	0.151	0.179	0.197
80	0.098	0.126	0.150	0.177	0.196
81	0.098	0.125	0.149	0.176	0.195
82	0.097	0.124	0.148	0.175	0.194
83	0.097	0.124	0.147	0.174	0.192
84	0.096	0.123	0.146	0.173	0.191
85	0.095	0.122	0.145	0.172	0.190
86	0.095	0.121	0.144	0.171	0.189
87	0.094	0.121	0.144	0.170	0.187
88	0.094	0.120	0.143	0.169	0.187
89	0.093	0.119	0.141	0.168	0.185
90	0.093	0.119	0.141	0.167	0.185
91	0.092	0.117	0.140	0.166	0.183
92	0.091	0.117	0.139	0.165	0.182
93	0.091	0.116	0.138	0.164	0.181
94	0.090	0.116	0.137	0.163	0.180
95	0.090	0.115	0.137	0.162	0.179
96	0.089	0.114	0.136	0.161	0.179
97	0.089	0.114	0.135	0.160	0.177
98	0.089	0.113	0.135	0.160	0.177
99	0.088	0.113	0.134	0.159	0.175
100	0.087	0.112	0.133	0.158	0.1