Calcul du tau de Kendall avec une variable aléatoire discrète et une continue.

Sous la supervision de
Étienne Marceau

Rapport des travaux réalisés

Préparé par

Alexandre Lepage,
Diamilatou N'diaye,

le 18 juin 2019

Faculté des sciences et de génie
École d'actuariat
Université Laval

1 Motivation

Dans la littérature, on explique comment calculer les tau de Kendall avec des variables aléatoire continues ([Genest and Favre, 2007]) et avec des variables aléatoires discrètes ([Nikoloulopoulos and Karlis, 2010]). Mais qu'en est-il si on a une variable aléatoire discrète et que l'autre est continue. Le lemme 1 définit la façon de calculer ce cas particulier.

2 Calcul du tau de Kendall avec une v.a. discrète et une v.a. continue.

Soit C, une copule quelconque, et les variables aléatoires N et X définies sur $\mathbb{N}$ et $\mathbb{R}_+$ respectivement. Avec le théorème de Sklar, on a que $F_{N,X}(n,x) = C(F_N(n), F_X(x)))$ et la fonction de densité bivariée conjointe est donnée par

$$f_{N,X}(n,x) = \frac{\partial}{\partial x} C(F_N(n), F_X(x))) - \frac{\partial}{\partial x} C(F_N(n-1), F_X(x))).$$

Soient les couples de variables aléatoires $\{(X_1, Y_1), \ldots, (X_k, Y_k)\}$. Un couple de v.a. est dit concordant si $(X_i - X_j)(Y_i - Y_j) > 0$ et discordant si $(X_i - X_j)(Y_i - Y_j) < 0$, pour $i \neq j$. Dans [Nikoloulopoulos and Karlis, 2010], on pose qu'en cas de possibilité d'égalité entre des observations d'une même variable aléatoire, la formule générale du tau de Kendall est

$$\begin{aligned}
\tau(N,X) &= \mathbb{P}(Concordance) - \mathbb{P}(Discordance) \\
&= \mathbb{P}(Concordance) - \big[1 - \mathbb{P}(Concordance) - \mathbb{P}(\acute{E}galit\acute{e})\big] \\
&= 2\,\mathbb{P}(Concordance) + \mathbb{P}(\acute{E}galit\acute{e}) - 1. \qquad (1)
\end{aligned}$$

Lemme 1. *Soient les couples de variables aléatoires $\{(N_1, X_1), \ldots, (N_k, X_k)\}$ définis sur $\mathbb{N}^k \times \mathbb{R}_+^k$. La formule pour calculer le tau de Kendall avec une variable aléatoire discrète et une autre qui est continue est présentée en* (2).

$$\tau(N,X) = 4 \sum_{n=0}^{\infty} \int_0^{\infty} F_{N,X}(n-1,x) f_{N,X}(n,x)\,dx + \sum_{n=0}^{\infty} \left(\mathbb{P}(N=n)\right)^2 - 1. \qquad (2)$$

Démonstration. De (1), on a

$$\tau(N,X) = 4\,\mathbb{P}(N_i < N_j, X_i < X_j) + \mathbb{P}(N_i = N_j \cup X_i = X_j) - 1,\ i \neq j. \qquad (3)$$

Or

$$\begin{aligned}
\mathbb{P}(N_i < N_j, X_i < X_j) &= \sum_{n=0}^{\infty} \int_{x=0}^{\infty} \mathbb{P}(N_i < n, X_i < x \cap N_j = n, X_j = x)\mathrm{d}x \\
&\overset{\text{ind.}}{=} \sum_{n=0}^{\infty} \int_{x=0}^{\infty} \mathbb{P}(N_i < n, X_i < x)\,\mathbb{P}(N_j = n, X_j = x)\mathrm{d}x \\
&= \sum_{n=0}^{\infty} \int_{x=0}^{\infty} \mathbb{P}(N_i \leq n-1, X_i < x)\, f_{N,X}(n,x)\mathrm{d}x \\
&= \sum_{n=0}^{\infty} \int_{x=0}^{\infty} F_{N,X}(n-1,x) f_{N,X}(n,x)\mathrm{d}x. \qquad (4)
\end{aligned}$$

Par la suite, puisque X est une v.a. continue,

$$\mathbb{P}(N_i = N_j \cup X_i = X_j) = \mathbb{P}(N_i = N_j).$$

On obtient alors

$$\begin{aligned}\mathbb{P}(N_i = N_j \cup X_i = X_j) &= \sum_{n=0}^{\infty} \mathbb{P}(N_i = n \cap N_j = n) \\ &\overset{\text{ind.}}{=} \sum_{n=0}^{\infty} \mathbb{P}(N_i = n)\, \mathbb{P}(N_j = n) \\ &\overset{\text{i.d.}}{=} \sum_{n=0}^{\infty} \left(\mathbb{P}(N = n)\right)^2. \end{aligned} \tag{5}$$

En insérant (4) et (5) dans (3), on obtient

$$\tau(N, X) = 4 \sum_{n=0}^{\infty} \int_0^{\infty} F_{N,X}(n-1, x) f_{N,X}(n, x) \mathrm{d}x + \sum_{n=0}^{\infty} \left(\mathbb{P}(N = n)\right)^2 - 1.$$

□

3 Calcul du tau de Kendall empiriquement.

Pour ce qui est du calcul empirique de deux variables aléatoires continues, [Genest and Favre, 2007] propose (6).

$$\begin{aligned}\tau_n(X, Y) &= \frac{P_n - Q_n}{\binom{n}{2}} \\ &= \frac{P_n - (\binom{n}{2} - P_n)}{\binom{n}{2}} \\ &= \frac{2\,P_n - \binom{n}{2}}{\binom{n}{2}} \\ &= \frac{4\,P_n}{n(n-1)} - 1, \end{aligned} \tag{6}$$

où P_n et Q_n représentent les nombres de couples concordants et discordants respectivement et n est le nombre d'observations.

Lemme 2. *Dans le cas où on a au moins une variable aléatoire discrète, dans la séquence de couples* $\{(N_1, X_1), \ldots, (N_k, X_k)\}$, (6) *devient*

$$\tau_n(N, X) = \frac{4\,P_n + 2\,E_n}{n(n-1)} - 1, \tag{7}$$

où E_n *est le nombre d'observations avec au moins une égalité.*

Démonstration. De façon similaire à (6), dans le cas où on a au moins une variable aléatoire discrète, on a

$$\begin{aligned}\tau_n(N, X) &= \frac{P_n - Q_n}{\binom{n}{2}} \\ &= \frac{P_n - (\binom{n}{2} - P_n - E_n)}{\binom{n}{2}} \\ &= \frac{2\,P_n + E_n - \binom{n}{2}}{\binom{n}{2}} \\ &= \frac{2\,P_n + E_n}{\binom{n}{2}} - 1 \\ &= \frac{4\,P_n + 2\,E_n}{n(n-1)} - 1. \end{aligned} \tag{8}$$

□

3.1 Calcul numérique

Du point de vue computationnel, (7) peut peut être écrit en `R` comme dans le code 1

Code 1.

```
tau_kendall <- function(X,Y){
    n <- length(X)
    concord <- outer(1:n,1:n, function(i,j) sign((X[i] - X[j]) * (Y[i] - Y[j])))
    E_n <- (length(concord[concord == 0]) - n) / 2
    P_n <- sum(concord[concord > 0]) / 2
    tau <- (4 * P_n + 2 * E_n) / (n * (n - 1)) - 1
    return(tau)
}
```

Il se trouve que la fonction `R cor` donne des résultats similaires, mais pas exactement les mêmes comme on peut le voir dans l'exemple 1.

Exemple 1. *Soient 20 réalisations du couple (X_i, Y_i) simulées à l'aide de la copule FGM dont le paramètre de dépendance est de 0.5. On a que $X_i \sim X \sim Y_i \sim Y \sim Poisson(10)$.*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
X	9	7	7	14	9	10	11	11	8	10	6	16	10	12	7	9	12	7	6	13
Y	17	10	6	7	9	7	8	11	4	10	12	14	10	10	9	12	4	7	10	7

Tableau 1 – Réalisations du couple (X, Y), où $X \sim Y \sim Poisson(10)$.

Fonction empirique	-0.0579
Fonction cor de R	-0.0636
Tau théorique (continue)	0.1111

Tableau 2 – Comparaison des résultats obtenus avec de la fonction présentée dans le code 1, la fonction `R cor` et le τ théorique qui est obtenu dans le cas continu.

Dans le tableau 2, on voit d'abord la nette différence entre le tau de Kendall théorique (continu) et celui calculé avec le code 1 ou avec la fonction `R cor`. Cela s'explique par le grand nombre d'égalités dans les données simulées, du fait que les deux variables sont discrètes et que le paramètre des lois de Poisson est relativement faible pour le nombre de simulations. Par la suite, on voit qu'il existe une différence entre le résultat du code 1 et de la fonction `cor`.

Références

[Genest and Favre, 2007] Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*.

[Nikoloulopoulos and Karlis, 2010] Nikoloulopoulos, A. K. and Karlis, D. (2010). Regression in a copula model for bivariate count data. *Journal of Applied Statistics*, 37(9) :1555–1568.