UNIVERSITÉ DE QUÉBEC

MÉMOIRE PRÉSENTÉ À
UNIVERSITÉ DE QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN MATHÉMATIQUES ET
INFORMATIQUE APPLIQUÉES

PAR
OUAIL ESSAADOUNI

QUELQUES GÉNÉRALISATIONS DES MESURES D'ASSOCIATION DE KENDALL ET DE SPEARMAN POUR DES DONNÉES DISCRÈTES MULTIVARIÉES

AUTOMNE 2007

RÉSUMÉ

Dans ce mémoire nous avons introduit et étudié des nouvelles mesures de dépendance multidimensionnelles appropriées aux données discrètes. Cette étude peut être vue comme une extension multivariée des résultats établis par Mesfioui et Tajar (2005). En particulier, nous avons établi la propriété de monotonicité de ces indices par rapport à l'ordre de concordance. Cette propriété a joué un rôle essentiel pour faciliter l'interprétation de ces mesures. Des versions empiriques de ces paramètres ont été examinées et illustrées via le modèle de Poisson multivarié. De nouveaux tests d'indépendance spécifiques aux données discrètes ont été établis.

REMERCIEMENTS

Ce mémoire à été réalisé sous la direction du Professeur Mhamed Mesfioui et M. Jean-François Quessy en co-direction. Je les remercie sincèrement d'avoir accepté de superviser ce travail et de m'avoir encouragé et soutenu durant toutes ces années.

Au professeur Mesfioui, un merci tout spécial pour ses conseils et son énorme appui tout au long de mon cursus universitaire. Au professeur Quessy, toute ma gratitude pour son aide technique chaleureusement accordée et à son sens de la minutie qu'il m'a inculqué.

J'exprime ici ma très vive reconnaissance envers mes chers parents et ma sœur Ouiam pour leur dévouement et leur soutien dans les moments difficiles.

Une partie des mes travaux a été financé par des octrois consentis aux professeurs Mesfioui et Quessy par le Conseil de recherche en sciences naturelles et en génie du Canada. Je tiens à leur exprimer ma gratitude pour cet appui financier.

Table des matières

LISTE DES TABLEAUX

LISTE DES FIGURES

CHAPITRE 1

INTRODUCTION

Une part importante des recherches effectuées sur les mesures de dépendance entre des variables aléatoires concerne le cas où ces variables sont continues. Cependant, beaucoup de phénomènes à étudier font intervenir des variables discrètes, pour lesquelles les valeurs observées peuvent prendre seulement un nombre dénombrable de valeurs. De nombreux domaines, notamment les sciences humaines et les sciences des affaires, font appel à des données discrètes.

Un exemple simple de variable discrète survient lorsque seulement les valeurs 0 et 1 sont observées. Souvent, 0 représente l'*absence* et 1 la *présence* d'un certain caractère, comme être en faveur ou en défaveur d'une nouvelle politique gouvernementale. On peut également penser à une épreuve où survient soit un succès, représenté par 1, ou un échec, dénoté 0.

Une variable peut également représenter le nombre d'événements survenus pendant une période de temps donné. Dans ce cas, l'ensemble des valeurs possibles est $\{0, 1, 2, \ldots, \}$. Par exemple, en assurance, on pourrait s'intéresser

à compter le nombre de catastrophes naturelles qui se sont produites lors des dix dernières années.

Dans ce travail, une attention particulière sera portée à la modélisation simultanée de plusieurs variables de nature discrète. Une façon populaire de quantifier la dépendance entre deux variables consiste à considérer une *mesure de dépendance.* La plus utilisée est le coefficient de corrélation linéaire de Pearson, mais les conditions d'application restrictives de cette mesure, notamment que les lois marginales des variables considérées soient normales, ont incité plusieurs chercheurs à proposer des alternatives pour évaluer la dépendance. À ce titre, le tau de Kendall et le rho de Spearman, dont l'utilisation est valide peu importe la loi des observations, sont les plus utilisés. Les propriétés de ces statistiques sont bien connues dans le cas continue.

Ce mémoire est structuré comme suit. Au chapitre 2, nous rappellerons les définitions du tau de Kendall et du rho de Spearman dans les cas continu. Au chapitre 3 nous allons présenter des extensions des définitions de ces mesures au cas discret, une illustration sera examinée via le modèle de Poisson bivarié. Le chapitre 4 fera l'objet d'une extension multivariée du tau de Kendall et du rho de Spearman discrets. En particulier, nous montrerons que la propriété de monotonicité reste valable pour ces nouvelles mesures de dépendance. Le chapitre 5 sera consacré à une étude statistique de ces paramètres. Des tests d'indépendance basés sur le tau de Kendall discret seront établies et illustrés par le modèle de Poisson multidimensionnelle.

CHAPITRE 2

MESURES ET CONCEPTS DE DÉPENDANCE

2.1 Théorie des copules

Soient X_1 et X_2, des variables aléatoires de fonction de répartition jointe

$$H(x_1, x_2) = \mathrm{P}\left(X_1 \leq x_1, X_2 \leq x_2\right)$$

et de fonctions de répartition marginales

$$F_1(x) = \mathrm{P}(X_1 \leq x) = H(x, \infty) \quad \text{et} \quad F_2(x_2) = \mathrm{P}(X_2 \leq y) = H(\infty, x).$$

La fonction de survie associée à H est définie par $\bar{H}(x, y) = \mathrm{P}\left(X_1 > x_1, X_2 > x_2\right)$. Puisque si A et B sont deux événements, alors $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$, on déduit que

$$\begin{aligned} \bar{H}(x_1, x_2) &= 1 - \mathrm{P}\left(X_1 \leq x_1 \cup X_2 \leq x_2\right) \\ &= 1 - \{\mathrm{P}(X_1 \leq x_1) + \mathrm{P}(X_2 \leq x_2) - \mathrm{P}(X_1 \leq x_1, X_2 \leq x_2)\} \\ &= 1 - F_1(x_1) - F_2(x_2) + H(x_1, x_2). \end{aligned}$$

Il est possible d'établir des bornes entre lesquelles H se situe. En effet, comme $A \cap B \subseteq A$, on a toujours que $\mathrm{P}(A \cap B) \leq \mathrm{P}(A)$. En appliquant ce résultat à H, on déduit que

$$H(x_1, x_2) \leq \mathrm{P}(X_1 \leq x_1) = F_1(x_1) \quad \text{et} \quad H(x_1, x_2) \leq \mathrm{P}(X_2 \leq x_2) = F_2(x_2).$$

Ainsi, $H(x_1, x_2) \leq \min\{F_1(x_1), F_2(x_2)\}$. D'autre part, puisque H est une probabilité, il est assuré que $H(x_1, x_2) \geq 0$ pour tout x_1, x_2 De même, $\bar{H}(x, y)$ est aussi une probabilité, et ainsi

$$0 \leq \bar{H}(x, y) = 1 - F_1(x_1) - F_2(x_2) + H(x_1, x_2).$$

On en tire l'inégalité

$$H(x_1, x_2) \geq \max\{F_1(x_1) + F_2(x_2) - 1,\ 0\}.$$

Pour tout x, y, la chaîne d'inégalités

$$\max\{F_1(x_1) + F_2(x_2) - 1,\ 0\} \leq H(x_1, x_2) \leq \min\{F_1(x_1), F_2(x_2)\} \tag{2.1}$$

est donc toujours vraie, peu importe la fonction de répartition H. Il est important de noter que les fonctions qui bornent H dans l'équation (2.1) sont elles-mêmes des fonctions de répartition bivariées, ce qui veut dire que ces bornes ne peuvent pas être améliorées.

Un résultat important de Sklar (1959) [1] établit que si les marges F et G de H sont continues, alors il existe une unique fonction $C : [0, 1]^2 \to [0, 1]$ telle que pour tout $(x_1, x_2) \in \mathbb{R}^2$, on a

$$H(x_1, x_2) = C\{F_1(x_1), F_2(x_2)\}. \tag{2.2}$$

La fonction C s'appelle une *copule*. Ce résultat permet donc d'établir une relation, à l'aide d'une fonction multivariée qui prend ses valeurs uniquement sur le carré unitaire $[0,1]^2$, entre la loi jointe d'une distribution et ses marges. Ainsi, C contient toute l'information à propos de la dépendance qui existe entre les variables aléatoires X et Y.

Le résultat de Sklar peut également s'utiliser à l'inverse, c'est-à-dire qu'on peut extraire la copule associée à une certaine fonction de répartition bivariée H de marges continues F et G. Pour ce faire, posons $u_1 = F_1(x_1)$ et $u_2 = F_2(x_2)$. Ainsi, de l'équation (2.2), on déduit que l'unique copule associée à H est

$$C(u_1, u_2) = H\left\{F_1^{-1}(u_1), F_2^{-1}(u_2)\right\}. \tag{2.3}$$

En particulier, ce résultat permet de réécrire les bornes (2.1) en termes d'une copule C. On déduit en effet que pour toute copule C,

$$W(u_1, u_2) \leq C(u_1, u_2) \leq M(u_1, u_2),$$

où W et M sont respectivement les bornes inférieure et supérieure de Fréchet–Hoeffding définies par

$$W(u_1, u_2) = \max\left(u_1 + u_2 - 1,\ 0\right) \quad \text{et} \quad M(u_1, u_2) = \min(u_1, u_2).$$

Une autre copule importante est celle associée à l'indépendance entre des variables aléatoires. On sait que dans ce cas, la fonction de répartition conjointe s'écrit $H(x_1, x_2) = F_1(x_1)F_2(x_2)$, ce qui implique, de l'équation (2.3), que la copule d'indépendance est

$$\Pi(u_1, u_2) = H\left\{F_1^{-1}(u_1), F_2^{-1}(u_2)\right\} = u_1 u_2.$$

Exemple 2.1. *Pour illustrer l'application du Théorème de Sklar, soient deux variables aléatoires X et X_2 de fonction de répartition*

$$H(x_1, x_2) = \left(1 - e^{-x_1} - e^{-x_2}\right)^{-1}, \quad x_1 \geq 0,\ x_2 \geq 0.$$

Les marges de H_1 sont

$$F_1(x_1) = \left(1 - e^{-x_1}\right)^{-1} \quad et \quad F_2(x_2) = \left(1 - e^{-x_2}\right)^{-1},$$

de telle sorte que

$$F_1^{-1}(u) = F_2^{-1}(u) = -\log\left(1 - \frac{1}{u}\right).$$

En appliquant l'équation (2.3)*, on déduit l'unique copule C_1 associée à H_1, à savoir*

$$\begin{aligned} C(u_1, u_2) &= \left\{1 - e^{-F_1^{-1}(u_1)} - e^{-F_2^{-1}(u_2)}\right\}^{-1} \\ &= \left\{1 - \left(1 - \frac{1}{u_1}\right) - \left(1 - \frac{1}{u_2}\right)\right\}^{-1} \\ &= \left(\frac{1}{u_1} + \frac{1}{u_2} - 1\right)^{-1} \\ &= \frac{u_1 u_2}{u_1 + u_2 - u_1 u_2}. \end{aligned}$$

Ces résultats s'étendent facilement au cas multivarié. En effet, la plupart des théorèmes et définitions obtenus dans le cas bivarié ont des versions analogues pour $d > 2$ variables.

Soient $X_1, \ldots, X_d$ des variables aléatoires de fonction de répartition jointe

$$H(x_1, \ldots, x_d) = P(X_1 \leq x_1, \ldots, X_d \leq x_d)$$

et de fonctions de répartition marginales

$$F_i(x) = \mathrm{P}(X_i \leq x), \quad 1 \leq i \leq d.$$

L'analogue multivarié du Théorème de Sklar (1959) implique que si les marges F_i sont continues, alors il existe une unique fonction $C : [0,1]^d \to [0,1]$ telle que pour tout $(x_1, \ldots, x_d) \in \mathbb{R}^d$, on a

$$H(x_1, \ldots, x_d) = C\left\{F_1(x_1), \ldots, F_n(x_d)\right\}. \tag{2.4}$$

Comme dans le cas $d = 2$, on peut utiliser le résultat de Sklar à l'inverse pour extraire la copule correspondante à une certaine fonction de répartition multivariée H de marges continues $F_1, \ldots, F_d$. Pour se faire, posons $u_i = F(x_i)$, $1 \leq i \leq d$ dans l'equation (2.4), on déduit que l'unique copule associée à H est

$$C(u_1, \ldots, u_d) = H\left\{F_1^{-1}(u_1), \ldots, F_n^{-1}(u_d)\right\}. \tag{2.5}$$

La copule d'indépendance multivariée est donnée par $\Pi^d(u_1, \ldots, u_d) = \prod_{i=1}^d u_i$. Les extensions pour $d > 2$ des bornes de Fréchet–Hoeffding, c'est-à-dire les versions multivariées de M et W, sont

$$\begin{aligned} M^d(u_1, \ldots, u_d) &= \min(u_1, \ldots, u_d) \\ W^d(u_1, \ldots, u_d) &= \max(u_1 + \cdots + u_d - d + 1,\ 0) \end{aligned}$$

À noter que M^d et Π^d sont des copules pour tout $d \geq 2$. Cependant, W^d n'est pas une copule pour $d > 2$. Ceci n'empêche pas d'avoir l'inégalité

$$W^d(u_1, \ldots, u_d) \leq C(u_1, \ldots, u_d) \leq M^d(u_1, \ldots, u_d),$$

valide pour toute copule d-variée C.

2.2 Mesures de dépendance bivariées

Le concept de dépendance a été largement étudié en probabilité et statistique, notamment dans le cas de deux variables aléatoires. Dans ce cas, plusieurs mesures de dépendances ont été proposées. Certaines sont basées sur la notion de *concordance* et de *discordance*, comme le tau de Kendall. D'autres sont basées sur les rangs des observations, comme le rho de Spearman.

Scarsini (1984) [2] a proposé une série de propriétés souhaitables pour une mesure de dépendance bivariée.

Définition 2.1. *Une mesure numérique d'association $\kappa(X,Y)$ entre deux variables aléatoires X et Y dont la copule est C est une mesure de concordance si et seulement si elle satisfait les propriétés qui suivent.*

1. *$\kappa(X,Y)$ est définie pour chaque couple (X,Y) de variables aléatoires;*
2. *$-1 \leq \kappa(X,Y) \leq 1$, $\kappa(X,X) = 1$ et $\kappa(X,-X) = -1$;*
3. *$\kappa(X,Y) = \kappa(Y,X)$;*
4. *Si X et Y sont indépendantes, alors $\kappa(X,Y) = 0$;*
5. *$\kappa(-X,Y) = \kappa(X,-Y) = -\kappa(X,Y)$;*
6. *Si les copules respectives de (X_1,Y_1) et (X_2,Y_2) sont telles que $C_1 \prec C_2$, alors $\kappa(X_1,Y_1) \leq \kappa(X_2,Y_2)$;*
7. *Si (X_n,Y_n) est une suite de variables aléatoires continues de copule C_n, et si C_n converge vers C, alors $\lim_{n\to\infty} \kappa(X_n,Y_n) = \kappa(X,Y)$, où $(X,Y) \sim C$.*

2.2.1 Le tau de Kendall

Soient (X_1, X_2) et $(\tilde{X}_1, \tilde{X}_2)$, deux vecteurs aléatoires continus indépendants et distribués selon la même loi H de marges F_1 et F_2. Dans ce qui suit, nous travaillerons sans perte de généralité avec les variables $U_i = F_1(X_i)$ et $\tilde{U}_i = F_2(\tilde{X}_i)$, $i = 1, 2$. Ainsi, la loi commune de (U_1, U_2) et $(\tilde{U}_1, \tilde{U}_2)$ sera l'unique copule C qu'on peut extraire de H par une application du Théorème de Sklar.

On dit que ces deux couples sont *concordants* si $U_1 < \tilde{U}_1, U_2 < \tilde{U}_2$ ou $U_1 > \tilde{U}_1, U_2 > \tilde{U}_2$. Ceci est équivalent à $(U_1 - \tilde{U}_1)(U_2 - \tilde{U}_2) > 0$. Dans le cas contraire, c'est à dire quand $(U_1 - \tilde{U}_1)(U_2 - \tilde{U}_2) < 0$, on dit que la paire est *discordante.*

Le tau de Kendall bivarié associé à la copule C, noté τ_C, est une mesure d'association non paramétrique entre deux variables aléatoires basée sur la notion de concordance. Spécifiquement, τ_C est défini comme la différence entre la probabilité de concordance et la probabilité de discordance. On a

$$\begin{aligned}
\tau_C &= \mathrm{P}\left\{(U_1 - \tilde{U}_1)(U_2 - \tilde{U}_2) > 0\right\} - \mathrm{P}\left\{(U_1 - \tilde{U}_1)(U_2 - \tilde{U}_2) < 0\right\} \\
&= 2\mathrm{P}\left\{(U_1 - \tilde{U}_1)(U_2 - \tilde{U}_2) > 0\right\} - 1 \\
&= 2\left\{\mathrm{P}\left(U_1 < \tilde{U}_1, U_2 < \tilde{U}_2\right) + \mathrm{P}\left(U_1 > \tilde{U}_1, U_2 > \tilde{U}_2\right)\right\} - 1 \\
&= 4\mathrm{P}\left(U_1 < \tilde{U}_1, U_2 < \tilde{U}_2\right) - 1 \\
&= 4\int_0^1 \int_0^1 \mathrm{P}\left(U_1 < u_1, U_2 < u_2 \mid \tilde{U}_1 = u_1, \tilde{U}_2 = u_2\right) \mathrm{d}C(u, v) \\
&= 4\int_0^1 \int_0^1 C(u_1, u_2)\, \mathrm{d}C(u_1, u_2) - 1.
\end{aligned}$$

Il est possible d'estimer τ_C avec un échantillon de données bivariées

$$(X_1, Y_1), \ldots, (X_n, Y_n).$$

En effet, si $\mathcal{B}_n$ dénote le nombre de concordances parmi les $n(n-1)/2$ comparaisons de paires possibles, c'est à dire

$$\mathcal{B}_n = \sum_{i<j} \mathbf{1}\left\{(X_{1i} - X_{ij})\,(X_{2i} - X_{2j}) > 0\right\}$$

alors le tau de Kendall empirique s'écrit

$$\tau_n = \frac{4\mathcal{B}}{n(n-1)} - 1.$$

Pour la borne inférieure de Fréchet–Hoeffding, le tau de Kendall vaut

$$\tau_W = 4\int_0^1 W(u, 1-u)\,\mathrm{d}u - 1 = -1,$$

car toute la masse de probabilité de $\mathrm{d}W(u,v)$ se retrouve sur la droite $v = 1-u$. Pour la copule d'indépendance, on a

$$\tau_\Pi = 4\int_0^1\int_0^1 uv\,\mathrm{d}u\,\mathrm{d}v - 1 = 4\left(\int_0^1 u\,\mathrm{d}u\right)^2 - 1 = 4(1/4) - 1 = 0,$$

alors que la borne supérieure de Fréchet–Hoeffding amène

$$\tau_M = 4\int_0^1 M(u,u)\,\mathrm{d}u - 1 = 4\int_0^1 u\,\mathrm{d}u - 1 = 4(1/2) - 1 = 1,$$

car toute la masse de probabilité de $\mathrm{d}M(u,v)$ est concentrée en $u = v$.

2.2.2 Le rho de Spearman

Prenons maintenant trois vecteurs aléatoires continus (X_1, X_2), $(\tilde{X}_1, \tilde{X}_2)$ et (X_1', X_3'), distribués selon la même loi H de marges F_1 et F_2. À l'instar

des arguments précédents, on travaillera sans perte de généralité avec les variables $U_i = F_i(X_i)$, $\tilde{U}_i = F_i(\tilde{X}_i)$ et $U_i' = F_i(X_i')$, $i = 1, 2$. Ainsi, la loi commune de (U_1, U_2), $(\tilde{U}_1, \tilde{U}_2)$ et (U_1', U_2') sera l'unique copule C qu'on peut extraire de H par une application du Théorème de Sklar.

Le rho de Spearman, aussi appelé le coefficient de corrélation de rangs, est une mesure d'association introduite par Spearman (1904). Tout comme le tau de Kendall, la valeur du rho de Spearman pour une loi H donnée ne dépend pas des marges, mais uniquement de la copule C associé à H. Sa définition théorique est

$$\rho_C = 3\mathrm{P}\left\{(U1 - \tilde{U}_1)(U_2 - U_2') > 0\right\} - 3\mathrm{P}\left\{(U_1 - \tilde{U}_1)(U_2 - U_3') < 0\right\}. \quad (2.6)$$

On peut exprimer ρ_C par une fonctionnelle qui dépend uniquement de C. D'abord,

$$\begin{aligned}
&\mathrm{P}\left\{(U_1 - \tilde{U}_1)(U_2 - U_2') > 0\right\} \\
&\quad = \int_0^1 \int_0^1 \mathrm{P}\left\{(U_1 - \tilde{U}_1)(U_2 - U_2') > 0 \mid \tilde{U}_1 = u_1, U_2' = u_2\right\} \mathrm{d}u\,\mathrm{d}v \\
&\quad = \int_0^1 \int_0^1 \left\{\mathrm{P}\left(U_1 > u_1, U_2 > u_2\right) + \mathrm{P}\left(U_1 < u_1, U_2 < u_2\right)\right\} \mathrm{d}u_1\,\mathrm{d}u_2 \\
&\quad = \int_0^1 \int_0^1 \left\{1 - u_1 - u_2 + 2C(u_1, u_2)\right\} \mathrm{d}u_1\,\mathrm{d}u_2 \\
&\quad = 2\int_0^1 \int_0^1 C(u_1, u_2)\,\mathrm{d}u_1\,\mathrm{d}u_2.
\end{aligned}$$

De même, du calcul précédent,

$$\begin{aligned}
\mathrm{P}\left\{(U_1 - \tilde{U}_1)(U_2 - U_2') < 0\right\} &= 1 - \mathrm{P}\left\{(U_1 - \tilde{U}_1)(U_2 - U_2') > 0\right\} \\
&= 1 - 2\int_0^1 \int_0^1 C(u_1, u_2)\,\mathrm{d}u_1\,\mathrm{d}u_2.
\end{aligned}$$

De là, on déduit que

$$\rho_C = 12 \int_0^1 \int_0^1 C(u_1, u_2)\, \mathrm{d}u_1\, \mathrm{d}u_2 - 3. \tag{2.7}$$

La version échantillonnale de ρ_C est

$$\hat{\rho}_n = \frac{12}{n(n^2-1)} \sum_{i=1}^{n} R_i S_i - 3\left(\frac{n+1}{n-1}\right),$$

où R_i est le rang de X_i parmi $X_1, \ldots, X_n$ et S_i est le rang de Y_i parmi $Y_1, \ldots, Y_n$.

2.3 Extension des mesures de dépendance au cas multivarié

2.3.1 Tau de Kendall d-varié : cas continu

Soit $X = (X_1, \ldots, X_d)$, un vecteur aléatoire continu à valeurs dans $\mathbb{R}^n$, de fonction de répartition jointe H, de copule associée C, et de marges $F_1, \ldots, F_d$. Pour construire le tau de Kendall multivarié, on suppose que cette mesure s'écrit linéairement en fonction de $\mathrm{E}_H\{H(X)\}$, à l'instar du cas bivarié. Autrement dit, on pose

$$\tau_C^d(X) = a\mathrm{E}_H\{H(X)\} + b,$$

où a et b sont des constantes réelles dont la valeur sera déterminée par les restrictions R1–R2 qui suivent.

R1. $\tau_\Pi^d = 0$, où $\Pi^d(u_1, \ldots, u_d) = u_1 \ldots u_d$ est la copule d'indépendance;

R2. $\tau_M^d = 1$ où $M^d = (u_1, \ldots u_d)$ est la borne de Fréchet supérieure.

Dans le cas de variables indépendantes, $H(x) = F_1(x_1) \cdots F_d(x_d)$. Ainsi,

$$\begin{aligned} \mathrm{E}_H\{H(X)\} &= \mathrm{E}_H\{F_1(X_1) \cdots F_d(X_d)\} \\ &= \mathrm{E}\{F_1(X_1)\} \cdots \mathrm{E}\{F_d(X_d)\} \\ &= \frac{1}{2^d}, \end{aligned}$$

car $F_i(X_i)$ suit une loi uniforme sur $(0,1)$. Maintenant, lorsque $X \sim M^d$, on sait qu'il existe une variable V uniformément distribuée sur $(0,1)$ telle que $X_i = F_i^{-1}(V)$. Donc, puisqu'alors $H(X) = \min\{F_1(X_1), \ldots, F_d(X_d)\}$, on a

$$\begin{aligned} \mathrm{E}_H\{H(X_1, \ldots, X_d)\} &= \mathrm{E}\left\{H\left(F_1^{-1}(V), \ldots, F_d^{-1}(V)\right)\right\} \\ &= \mathrm{E}\left\{\min\left(F_1 \circ F_1^{-1}(V), \ldots, F_d \circ F_d^{-1}(V)\right)\right\} \\ &= \mathrm{E}(V) \\ &= 1/2. \end{aligned}$$

Ainsi en combinant les résultats précédents, on tire les équations

$$\frac{a}{2^d} + b = 0 \quad \text{et} \quad \frac{a}{2} + b = 1,$$

dont l'unique solution est

$$a = \frac{2^d}{2^{d-1} - 1} \quad \text{et} \quad b = \frac{-1}{2^{d-1} - 1}.$$

On conclue que le tau de Kendall dans le cas continu multivarié est donné par la formule

$$\tau_C^d(X) = \frac{2^d \mathrm{E}_H\{H(X)\} - 1}{2^{d-1} - 1} \tag{2.8}$$

À noter que cette formule multivariée généralise l'expression du tau de Kendall dans le cas continu bivarié, c'est-à-dire quand $d = 2$.

2.3.2 Rho de Spearman : cas continu

Pour construire le rho de Spearman dans un contexte multivarié, on suppose que ce paramètre est une combinaison linéaire de $\mathrm{E}_{\Pi}\{H(X)\}$, à l'instar du cas bivarié. Autrement dit, on pose

$$\rho_C^d(X) = a\mathrm{E}_{\Pi}\{H(X)\} + b,$$

où a et b sont choisies pour satisfaire les contraintes R1–R2. Selon le calcul effectué pour le tau de Kendall,

$$\mathrm{E}_H\{H(X)\} = \frac{1}{2^d}$$

quand les variables sont indépendantes. Maintenant, lorsque $X \sim M^d$, on a par l'identité de Hoeffding que

$$\begin{aligned}\mathrm{E}_{\Pi}\{H(X_1, X_2, \ldots, X_d)\} &= \mathrm{E}_{M^d}\{F_1(X_1)\cdots F_d(X_d)\} \\ &= \mathrm{E}_V\{V^d\} \\ &= \frac{1}{d+1}.\end{aligned}$$

Ainsi, en combinant ces résultats, on a les équations

$$\frac{a}{2^d} + b = 0 \quad \text{et} \quad \frac{a}{d+1} + b = 1.$$

La solution à ce système est

$$a = \frac{(d+1)2^d}{2^d - d - 1} \quad \text{et} \quad b = -\frac{d+1}{2^d - d - 1},$$

Ceci amène la définition suivante pour l'extension multivariée du rho de Spearman

Définition 2.2. *La version d-varié du rho de Spearman est donnée par*

$$\rho^d(X) = \frac{(d+1)2^d \mathrm{E}_\Pi \{H(X)\} - d - 1}{2^d - d - 1}. \tag{2.9}$$

Cette formule généralise au cas multivarié le rho de Spearman bivarié.

2.4 Notions d'ordre

Deux variables aléatoires X_1 et X_2 sont dites concordantes lorsque de grandes valeurs de X_1 sont associées avec de grandes valeurs de X_2. Beaucoup de chercheurs ont tenté de formaliser cette définition intuitive, notamment Tchen (1980) [3] et Yanagimoto & Okamoto (1969) [4].

Soient (X_1, X_2) et $(\tilde{X}_1, \tilde{X}_2)$, deux vecteurs aléatoires de fonction de répartition jointes respectives H et $\tilde{H}$. Pour établir une relation d'ordre adéquate entre ces deux vecteurs, on suppose que les lois marginales sont identiques, c'est-à-dire que $F_(t) = \tilde{F}_1(t)$ et $F_2(t) = \tilde{F}_2(t)$ pour tout $t \in \mathbb{R}$. En effet, les relations d'ordre se veulent une comparaison entre la force de la dépendance dans un vecteur, ce qui ne devrait pas être affecté par le choix des marges.

On dit du vecteur (X_1, X_2) qu'il est *plus concordant* que $(\tilde{X}_1, \tilde{X}_2)$ si pour tout $(s,t) \in \mathbb{R}^2$,

$$H(s,t) \leq \tilde{H}(s,t).$$

On écrit alors $(X_1, X_2) \prec_c (\tilde{X}_1, \tilde{X}_2)$. Comme les marges sont identiques, cette définition est en fait uniquement déterminée par la comparaison des

copules C et $\tilde{C}$ sous-jacentes à H et $\tilde{H}$. Ainsi, $(X_1, X_2) \prec_c (\tilde{X}_1, \tilde{X}_2)$ si

$$C(u_1, u_2) \leq \tilde{C}(u_1, u_2)$$

pour tout $(u_1, u_2) \in (0,1)^2$.

La propriété suivante concernant l'ordre de concordance est tirée de Müller et Scarsini (2000) [5]. Elle trouvera son importance pour montrer la monotonicité de tau de Kendall et du rho de Spearman. Avant de l'énoncer, la définition d'une fonction quasi-monotone est donnée.

Définition 2.3. *On dit qu'une fonction ϕ est quasi-monotone si et seulement si pour tout $x_1, \tilde{x}_1, x_2, \tilde{x}_2$ tels que $x_1 \leq \tilde{x}_1$ et $x_2 \leq \tilde{x}_2$,*

$$\phi(\tilde{x}_1, \tilde{x}_2) - \phi(\tilde{x}_1, x_2) - \phi(x_1, \tilde{x}_2) + \phi(x_1, x_2) \geq 0.$$

Théorème 2.1. *Si (X_1, X_2) et $(\tilde{X}_1, \tilde{X}_2)$ sont des vecteurs aléatoires de mêmes fonctions de répartition marginales, alors*

$$(X_1, X_2) \prec_c (\tilde{X}_1, \tilde{X}_2)$$

si et seulement si

$$\mathrm{E}\{\phi(X_1, X_2)\} \leq \mathrm{E}\left\{\phi(\tilde{X}_1, \tilde{X}_2)\right\}$$

pour toute fonction quasi-monotone ϕ.

La définition de l'ordre de concordance dans le cas de plusieurs variables est différente de celle dans le cas bivarié. Dans ce qui suit, la définition de Müller et Scarsini (2000) [5] sera adoptée.

Définition 2.4. *Soient* $X = (X_1, \ldots, X_d)$ *et* $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_d)$, *des vecteurs aléatoires de fonctions jointes respectives* H *et* $\tilde{H}$ *et de mêmes fonctions de répartition marginales* $F_1, \ldots, F_d$. *On note* $\bar{H}$ *et* $\bar{\tilde{H}}$ *les fonctions de survie associées, c'est-à-dire que* $\bar{H}(x_1, \ldots, x_d) = \mathrm{P}(X_1 > x_1, \ldots, X_d > x_d)$ *et* $\bar{\tilde{H}}(x_1, \ldots, x_d) = \mathrm{P}(\tilde{X}_1 > x_1, \ldots, \tilde{X}_d > x_d)$.

Définition 2.5. *On dit que* X *est plus petit que* $\tilde{X}$ *dans l'ordre quadrant inférieur noté* $X \leqslant_{lo} \tilde{X}$, *si les fonctions de répartition sont ordonnées, c'est-à-dire que pour tout* $(s_1, \ldots, s_d) \in \mathbb{R}^d$,

$$H(s_1, \ldots, s_d) \leq \tilde{H}(s_1, \ldots, s_d).$$

Définition 2.6. *On dit que* X *est plus petit que* $\tilde{X}$ *dans l'ordre quadrant supérieur noté* $X \leqslant_{uo} \tilde{X}$, *si les fonctions de survie sont ordonnées, c'est-à-dire que pour tout* $(s_1, \ldots, s_d) \in \mathbb{R}^d$,

$$\bar{H}(s_1, \ldots, s_d) \leq \bar{\tilde{H}}(s_1, \ldots, s_d).$$

Définition 2.7. *On dit que* X *est plus petit que* $\tilde{X}$ *dans l'ordre de concordance, noté* $X \leqslant_c \tilde{X}$, *si les définitions 1 et 2 sont satisfaites.*

Dans le cas bivarié, les définitions 1 et 2 précédentes sont équivalentes, c'est-à-dire que

$$H(s_1, s_2) \leq \tilde{H}(s_1, s_2)$$

implique que

$$\bar{H}(s_1, s_2) \leq \bar{\tilde{H}}(s_1, s_2)$$

pour tout $(s_1, s_2) \in \mathbb{R}^2$. En effet,

$$\begin{aligned} \bar{H}(s_1, s_2) &= 1 - F_1(s_1) - F_2(s_2) + H(s_1, s_2) \\ &\leq 1 - F_1(s_1) - F_2(s_2) + \tilde{H}(s_1, s_2) \\ &\leq \bar{\tilde{H}}(s_1, s_2). \end{aligned}$$

Ceci n'est pas vrai dès lors que $d > 2$, ce qui motive à adopter la définition précédente.

2.5 La dépendance positive

Soient X_1 et X_2, des variables aléatoires potentiellement dépendantes. On dit que X_2 est *décroissante du côté de la queue gauche* de X_1, qu'on note $LTD\,(X_2|X_2)$, si et seulement si

$$\mathrm{P}\,(X_2 \leq \tilde{x}_2 | X_1 \leq x_1)$$

est une fonction décroissante en x_1 pour toutes les valeurs possibles de x_1. De même, X_2 est dite *croissante du côté de la queue droite* de X_1, qu'on note $RTI\,(X_2|X_1)$, si et seulement si

$$\mathrm{P}\,(X_2 > x_2 | X_1 > x_1)$$

est une fonction croissante en x_1 pour toutes les valeurs possibles de x_2.

La croissance du côté de la queue gauche et la décroissance du côté de la queue droite, notées LTI et RTD, sont définies en échangeant les mots *croissante*

et décroissante dans les définitions ci-dessus.

Selon la proposition 2.3 de Capéraà et Genest (1993) [6], si X_1 et X_2 sont continues, alors le fait qu'on ait $RTI\,(X_2|X_1)$ et $LTD\,(X_2|X_1)$ implique que

$$\rho_C \geq \tau_C \geq 0$$

où τ_C est ρ_C sont le tau de Kendall et le rho de Spearman du couple (X_1, X_2) dont la copule sous-jacente à la loi jointe est C.

Dans le chapitre suivant, des versions de τ_C et ρ_C dans le cas des variables aléatoires discrètes seront proposées et une extension discrète du résultat précédent sera exposée.

CHAPITRE 3

MESURES DE DÉPENDANCE ENTRE DEUX VARIABLES DISCRÈTES

3.1 Contexte et illustrations

Ce chapitre est consacré à l'étude de quelques mesures de dépendance dans le cas où les variables d'intérêt sont discrètes. On s'intéressera particulièrement au tau de Kendall et au rho de Spearman. Certains résultats établis dans le cas continu ne seront plus valables dans le cas discret. Par exemple, les valeurs possibles du tau de Kendall et du rho de Spearman dans le cas discret ne seront plus nécessairement comprises entre -1 et 1. Ainsi, l'identification de variables parfaitement associées en tenant compte du degré d'association ne sera plus directement application dans le cas discret. En effet, il se peut que lors d'une association positive parfaite, le tau de Kendall soit égal à $1/2$. La fin de ce chapitre sera consacrée à l'étude de la monotonicité du tau de Kendall et du rho de Spearman par rapport à l'ordre de concordance. Cette propriété est ensuite illustrée par un modèle de Poisson bivarié. Des

simulations utilisant les versions empiriques de ces paramètres sont aussi présentées.

3.2 Tau de Kendall et rho de Spearman discrets

Soient deux vecteurs aléatoires indépendants (X, Y) et (X', Y'), chacun distribué selon la loi H. Alors en utilisant l'indépendance entre ces vecteurs et en conditionnant par rapport aux valeurs possibles de (X', Y'), on a

$$\mathrm{P}(X \leq X', Y \leq Y') = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} H(x, y) h(x, y) = \mathrm{E}\{H(X, Y)\}.$$

En particulier,

$$\mathrm{P}(X \leq X') = \mathrm{E}\{F(X)\} \quad \text{et} \quad \mathrm{P}(Y \leq Y') = \mathrm{E}\{G(Y)\}.$$

Contrairement au cas continu, il se peut que $\mathrm{P}(X = x) > 0$ pour une variable X discrète. Il s'ensuit alors que $\mathrm{P}(X \leq x) \neq \mathrm{P}(X < x)$. Pour tenir compte adéquatement de cette possibilité, définissons $H(x^-, y^-) = \mathrm{P}(X < x, Y < y)$, $H(x^-, y) = \mathrm{P}(X < x, Y \leq y)$ et $H(x, y^-) = \mathrm{P}(X \leq x, Y < y)$.

Le Lemme suivant sera utilisé fréquemment dans la suite.

Lemme 3.1. *Soit (X, Y), un couple de variables aléatoires de fonction de répartition jointe H et de distributions marginales F et G. Alors*

$$\mathrm{E}\{H(X, Y)\} = \mathrm{E}\left\{H\left(X^-, Y^-\right)\right\} - \mathrm{E}\left\{F\left(X^-\right)\right\} - \mathrm{E}\left\{G\left(Y^-\right)\right\} + 1.$$

Démonstration. On a

$$\begin{aligned}
\mathrm{E}\{H(X,Y)\} &= \mathrm{P}(X \leq X', Y \leq Y') \\
&= \mathrm{P}(X' \geq X, Y' \geq Y) \\
&= 1 - \mathrm{P}(X' < X) - \mathrm{P}(Y' < Y) + \mathrm{P}(X' < X, Y' < Y) \\
&= 1 - \mathrm{E}\left\{F(X^-)\right\} - \mathrm{E}\left\{G(Y^-)\right\} + \mathrm{E}\left\{H(X^-, Y^-)\right\}.
\end{aligned}$$

◊

Le résultat précédent permet d'obtenir une expression pour le tau de Kendall entre deux variables aléatoires discrètes.

Proposition 3.1. *Soient X et Y, deux variables aléatoires de fonction de répartition jointe H et de marges F et G respectivement. Alors le tau de Kendall du couple (X, Y) est donné par*

$$\begin{aligned}
\tau &= \mathrm{E}\{H(X,Y)\} + \mathrm{E}\left\{H(X^-, Y^-)\right\} \\
&\quad + \mathrm{E}\left\{H(X^-, Y)\right\} + \mathrm{E}\left\{H(X, Y^-)\right\} - 1. \qquad (3.1)
\end{aligned}$$

Démonstration. Par définition, le tau de Kendall est la différence entre la probabilité de concordance et la probabilité de discordance entre deux couples indépendants de même loi jointe, c'est-à-dire que

$$\tau = \mathrm{P}\left\{(X - X')(Y - Y') > 0\right\} - \mathrm{P}\left\{(X - X')(Y - Y') < 0\right\}.$$

Pour le premier terme à droite, c'est-à-dire la probabilité de concordance, on peut écrire

$$\begin{aligned}
\mathrm{P}\left\{(X - X')(Y - Y') > 0\right\} &= \mathrm{P}(X < X', Y < Y') + \mathrm{P}(X > X', Y > Y') \\
&= 2\mathrm{P}(X < X', Y < Y') \\
&= 2\mathrm{E}\left\{H(X^-, Y^-)\right\},
\end{aligned}$$

puisque (X,Y) et (X',Y') sont identiquement distribués. Pour le deuxième terme à droite, c'est-à-dire la probabilité de discordance, on a

$$\begin{aligned}
\mathrm{P}\left\{(X-X')(Y-Y')<0\right\} &= \mathrm{P}(X<X',Y>Y')+\mathrm{P}(X>X',Y<Y') \\
&= \mathrm{P}(X<X')-\mathrm{P}(X<X',Y\leq Y') \\
&\quad +\mathrm{P}(Y<Y')-\mathrm{P}(X\leq X',Y<Y') \\
&= \mathrm{E}\left\{F(X^-)\right\}-\mathrm{E}\left\{H(X^-,Y)\right\} \\
&\quad +\mathrm{E}\left\{G(Y^-)\right\}-\mathrm{E}\left\{H(X,Y^-)\right\}.
\end{aligned}$$

Puisque du Lemme (4.2), on déduit

$$\mathrm{E}\left\{F\left(X^-\right)\right\}+\mathrm{E}\left\{G\left(Y^-\right)\right\}=\mathrm{E}\left\{H\left(X^-,Y^-\right)\right\}-\mathrm{E}\left\{H(X,Y)\right\}+1,$$

on obtient l'expression annoncée pour τ. ◊

À noter que la formule du tau de Kendall dans le cas discret généralise l'expression déjà rencontrée pour des variables continues. En effet, si X et Y sont des variables aléatoires continues, alors

$$H(X^-,Y^-)=H(X^-,Y)=H(X,Y^-)=H(X,Y)$$

et $\tau = 4\mathrm{E}\left\{H(X,Y)\right\}-1 = 4\mathrm{E}\left\{C(U,V)\right\}-1$, où C est l'unique copule associée à H.

Proposition 3.2. *Soit (X,Y), un couple de variables aléatoires de fonction de répartition jointe H, et de fonctions de répartitions marginales F et G respectivement. Alors*

$$\begin{aligned}
\rho(X,Y) &= 3\mathrm{E}_\Pi\left\{H(X^-,Y)\right\}+3\mathrm{E}_\Pi\left\{H(X,Y^-)\right\} \\
&\quad +3\mathrm{E}_\Pi\left\{H(X^-,Y^-)\right\}+3\mathrm{E}_\Pi\left\{H(X,Y)\right\}-1.
\end{aligned}$$

Démonstration. Soient (X_1, Y_1), (X_2, Y_2) et (X_3, Y_3), des copies indépendantes de (X, Y), de telle sorte que

$$\rho(X,Y) = \mathrm{P}\left\{(X_1 - X_2)(Y_1 - Y_3) > 0\right\} - \mathrm{P}\left\{(X_1 - X_2)(Y_1 - Y_3) < 0\right\}.$$

On peut écrire cette expression sous la forme

$$\begin{aligned}
\rho(X,Y) &= 3\mathrm{P}(X_1 < X_2, Y_1 < Y_3) + 3\mathrm{P}(X_1 \leq X_2, Y_1 \leq Y_3] && (3.2)\\
&\quad + 3\mathrm{P}(X_1 < X_2, Y_1 \leq Y_3) + 3\mathrm{P}(X_1 \leq X_3, Y_1 < Y_3) && (3.3)\\
&\quad - 3\mathrm{P}(X_1 < X_2) + 3\mathrm{P}(X_1 > X_2) - 3\mathrm{P}(Y_1 < Y_3) + 3\mathrm{P}(Y_1 > Y_3) && (3.4)
\end{aligned}$$

Or, puisque (X_1, Y_1), (X_2, Y_2) et (X_3, Y_3) sont indépendants, on a

$$\mathrm{P}(X_1 < X_2, Y_1 < Y_3) = \mathrm{E}_\Pi\left\{H\left(X^-, Y^-\right)\right\}.$$

De la même façon,

$$\begin{aligned}
\mathrm{P}(X_1 \leq X_2, Y_1 \leq Y_3) &= \mathrm{E}_\Pi\left\{H(X,Y)\right\}\\
\mathrm{P}(X_1 < X_2, Y_1 \leq Y_3) &= \mathrm{E}_\Pi\left\{H\left(X^-, Y\right)\right\}\\
\mathrm{P}(X_1 \leq X_2, Y_1 < Y_3) &= \mathrm{E}_\Pi\left\{H\left(X, Y^-\right)\right\}.
\end{aligned}$$

Ainsi, en remplaçant les équations précédentes dans (3.2), on obtient les résultats envisagés. ◇

3.3 Monotonicité du tau de Kendall et du rho de Spearman discrets

Proposition 3.3. *Soient (X_1, Y_1) et (X_2, Y_2), deux couples de variables aléatoires de fonctions de répartition jointe respectives H_1 et H_2. Alors*

$$(X_1, Y_1) \prec_c (X_2, Y_2) \Rightarrow \tau_{H_1} \leq \tau_{H_2},$$

où τ_{H_i}, $i = 1, 2$, est le tau de Kendall d'une population de loi H_i.

Démonstration. Comme $H(x, y)$, $H_1(x^-, y^-)$, $H_1(x, y^-)$ et $H_1(x^-, y)$ sont des fonctions quasi-monotones, une application du théorème (2.1) amène $\mathrm{E}_{H_1}\left\{H_1\left(X_1^-, Y_1^-\right)\right\} \leq \mathrm{E}_{H_2}\left\{H_1\left(X_2^-, Y_2^-\right)\right\} \leq \mathrm{E}_{H_2}\left\{H_2\left(X_2^-, Y_2^-\right)\right\}$, où la dernière inégalité se déduit de $H_1(x, y) \leq H_2(x, y)$. Par des arguments identiques, on obtient également $\mathrm{E}_{H_1}\left\{H_1\left(X_1^-, Y_1\right)\right\} \leq \mathrm{E}_{H_2}\left\{H_2\left(X_2^-, Y_2\right)\right\}$ et $\mathrm{E}_{H_1}\left\{H_1\left(X_1, Y_1^-\right)\right\} \leq \mathrm{E}_{H_2}\left\{H_2\left(X_2, Y_2^-\right)\right\}$. De la définition du tau de Kendall discret, on obtient la conclusion annoncée. ◇

On obtient un résultat similaire pour la version discrète du rho de Spearman.

Proposition 3.4. *Soient (X_1, Y_1) et (X_2, Y_2) deux couples de variables aléatoires de fonctions de distributions jointe respectivement H_1 et H_2, alors*

$$(X_1, Y_1) \prec_c (X_2, Y_2) \Rightarrow \rho_{H_1} \leq \rho_{H_2},$$

où ρ_{H_i}, $i = 1, 2$, est le rho de Spearman d'une population de loi H_i.

3.4 Correction des versions discrètes de tau et rho

Dans le cas de variables continues X et Y, constater que $\tau = -1$ ou $\rho = -1$ signifie que X et Y sont parfaitement négativement dépendantes. De même, $\tau = 1$ ou $\rho = 1$ indique un lien positif parfait entre les deux variables. En fait, on a toujours $-1 \leq \tau \leq 1$ et $-1 \leq \rho \leq 1$.

Malheureusement, l'intervalle des valeurs possibles de τ et ρ est différent de $[-1, 1]$ dans le cas de variables discrètes. Ainsi, on a $\tau \in [\tau_{\min}, \tau_{\max}]$, où $\tau_{\min} > -1$ et $\tau_{\max} < 1$. De même, $\rho \in [\rho_{\min}, \rho_{\max}]$, où $\rho_{\min} > -1$ et $\rho_{\max} < 1$.

Une version modifiée du tau de Kendall est le coefficient tau_b proposé par Kendall(1945). La définition tient compte des probabilités d'égalités entre les composantes des vecteurs aléatoires, qui est non nulle dans le cas discret. La version théorique de ce coefficient de dépendance est

$$\tau_b = \frac{\tau}{\sqrt{\mathrm{P}\left(X_1 \neq X_2\right)\mathrm{P}\left(Y_1 \neq Y_2\right)}}.$$

Une autre mesure de dépendance a été proposée par Goodman et Kruskal (1954) [7]. Leur coefficient *gamma* est défini par

$$\gamma_H = \frac{\mathrm{P}_H(C) - \mathrm{P}_H(D)}{P_H(C) + P_H(D)}, \tag{3.5}$$

où $\mathrm{P}_H(C)$ est la probabilité de concordance pour deux couples (X, Y) et (X', Y') indépendants et de loi H, c'est-à-dire la probabilité de l'événement

$$C = \{X < X', Y < Y' \quad \text{ou} \quad X > X', Y > Y'\}.$$

La discordance D est quant à elle représentée par l'événement

$$D = \{X < X', Y > Y' \quad \text{ou} \quad X > X', Y < Y'\}.$$

La mesure γ_H constitue une correction intéressante pour τ car elle permet d'atteindre les valeurs extrêmes -1 et 1 pour la dépendance parfaite négative et positive, respectivement. Également, tel que le résultat suivant l'établit, γ_H est monotone, c'est-à-dire que $\gamma_H \leq \gamma_{H'}$ lorsque $H \leq H'$.

Proposition 3.5. *Soit le couple (X, Y) de loi H et le couple (X', Y') de loi H'. Si $H \leq H'$, alors $\gamma_H \leq \gamma_{H'}$.*

Démonstration. D'après l'équation (3.5), on peut écrire

$$\gamma_H = \frac{1 - \phi_H}{1 + \phi_H},$$

où $\phi_H = \mathrm{P}_H(D)/\mathrm{P}_H(C)$. Si $H \leq H'$, alors on déduit que

$$0 \leq P_{H'}(D) \leq P_H(D) \quad \text{et} \quad 0 \leq \mathrm{P}_H(C) \leq P_{H'}(C).$$

Par conséquent,

$$\phi_{H'} = \frac{\mathrm{P}_{H'}(D)}{\mathrm{P}_{H'}(C)} \leq \frac{\mathrm{P}_H(D)}{\mathrm{P}_H(C)} = \phi_H.$$

Comme γ_H est décroissante en fonction de ϕ_H, le fait que $\phi_{H'} \leq \phi_H$ assure que $\gamma_H \leq \gamma_{H'}$.

La correction proposée par Mesfioui, Tajar et Bouezmarni (2005) [9] consiste à transformer τ et ρ par

$$\tilde{\tau} = \begin{cases} \dfrac{\tau}{\tau_{\max}} & \text{si } \tau \geq 0 \\ -\dfrac{\tau}{\tau_{\min}} & \text{si } \tau < 0 \end{cases} \quad \text{et} \quad \tilde{\rho} = \begin{cases} \dfrac{\rho}{\rho_{\max}} & \text{si } \rho \geq 0 \\ -\dfrac{\rho}{\rho_{\min}} & \text{si } \rho < 0. \end{cases}$$

3.5 Continuation d'une variable discrète

Le principe de *continuation* consiste à transformer une variable discrète en variable continue. Pour illustrer, supposons que X_1 est une variable aléatoire discrète à valeurs dans $\mathbb{Z}$. Schriever (1985) [8] a proposé de transformer X_1 en variable continu par

$$X_1^\star = X_1 + U_1,$$

où U_1 est une variable aléatoire continue uniformément distribuée sur $[0, 1]$ et indépendante de X_1. On montre facilement que la fonction de répartition de X_1^* est

$$F_1^\star(x) = \sum_{i=-\infty}^{+\infty} F_U(x - i)\mathrm{P}(X_1 = i), \qquad x \in \mathbb{R},$$

où

$$F_U(u) = \begin{cases} 0, & u < 0 \\ u, & 0 \leq u \leq 1 \\ 1, & u > 1 \end{cases}$$

est la fonction de répartition de U.

Cette idée sera étendue ici au cas de deux variables discrètes X_1 et X_2 chacune à valeurs dans $\mathbb{Z}$ et dont la fonction de masse jointe est h. On propose les variables transformées

$$X_1^\star = X_1 + U_1 \quad \text{et} \quad X_2^\star = X_2 + U_2,$$

où U_1 et U_2 sont des variables uniformément distribuées sur $[0, 1]$ et indépendantes de X_1 et X_2. La fonction de répartition de $(X_1^\star, X_2^\star)$ s'écrit alors en fonction

de la copule C de (U_1, U_2) par

$$H^*(x_1, x_2) = \sum_{i=-\infty}^{+\infty} \sum_{j=-\infty}^{+\infty} C(x_1 - i, x_2 - j)\ h(i, j),$$

où $(x_1, x_2) \in \mathbb{R} \times \mathbb{R}$.

Soient maintenant $\tau_H^\star$ et $\rho_H^\star$, le tau de Kendall et le rho de Spearman du couple de variables continues $(X_1^\star, X_2^\star)$. Posons également τ_C et ρ_C comme étant les valeurs du tau de Kendall et du rho de Spearman pour le couple (U_1, U_2) de loi C. De la Proposition 3.1 de Mesfioui & Tajar (2005) [10], on a les relations

$$\tau_H^\star = \tau_H + \tau_C \mathrm{E}_H \left\{ S(X_1, X_2) \right\}$$

et

$$\rho_H^\star = \rho_H + \rho_C \mathrm{E}_\Pi \left\{ S(X_1, X_2) \right\},$$

où

$$S(X_1, X_2) = H(X_1, X_2) - H(X_1, X_2^-) - H(X_1^-, X_2) + H(X_1^-, X_2^-).$$

À remarquer que lorsque U_1 et U_2 sont indépendantes, alors $\tau_C = \rho_C = 0$ et par conséquent, $\tau_H^\star = \tau_H$ et $\rho_H^\star = \rho_H$.

L'utilité de la continuation est illustrée par la proposition suivante, qui permet d'étendre au cas discret quelques notions de dépendance présentées au Chapitre 2 pour des variables continues.

Proposition 3.6. *Soient X_1 et X_2, des variables aléatoires discrètes à valeurs dans $\mathbb{Z}$ dont la fonction de répartition jointe est H et les marges sont F_1 et*

F_2 respectivement. Soient les versions continues $X_i^\star = X_i + U_i$, où U_1 et U_2 sont indépendantes et uniformément distribuées sur $[0,1]$. Alors

$$LTD(X_2|X_1) \Longleftrightarrow LTD(X_2^\star|X_1^\star)$$

et

$$RTI(X_2|X_1) \Longleftrightarrow RTI(X_2^\star|X_1^\star).$$

Démonstration. Pour montrer la première équivalence, soit $(x_1, x_2) \in I_i \times I_j$, où $I_i = [i, i+1]$ et $i \in \mathbb{Z}$. On a

$$\mathrm{P}(X_2^\star \leq x_2 | X_1^\star > x_1) = \frac{(x_1 - i)\delta_1(x_2) + \delta_2(x_2)}{(x_1 - i)p_i + F_1(i-1)},$$

où

$$\delta_1(x) = (x - j)h(i,j) + H(i, j^-) - H(i^-, j^-)$$

et

$$\delta_2(x) = (x - j)\left\{H(i^-, j) - H(i^-, j^-)\right\} + H(i^-, j^-).$$

Par définition, $X_2^\star$ est *décroissante du côté de la queue gauche* si et seulement si pour tout $i, j \in \mathbb{Z}$, on a

$$\delta_1(x)F(i-1) - p_i\delta_2(x) \leq 0$$

pour tout $x \in I_j$. Si on définit

$$A_{ij} = h(i,j)F_1(i^-) - f_1(i)\left\{H(i^-, j) - H(i^-, j^-)\right\}$$

et

$$B_{ij} = F_1(i^-)\left\{H(i, j^-) - H(i^-, j^-)\right\} - f_1(i)H(i^-, j^-),$$

ceci est équivalent à

$$(x - j)A_{ij} + B_{ij} \leq 0$$

pour tout $x \in I_j$. Cette dernière inégalité est vraie si et seulement si, pour tout $i, j \in \mathbb{Z}$,

$$B_{ij} \leq 0 \quad \text{et} \quad A_{ij} + B_{ij} \leq 0.$$

Ces deux inégalités sont équivalentes à $LTD(X_2|X_1)$, c'est-à-dire que

$$\begin{aligned} B_{ij} \leq 0 &\iff F_1(i-1)H(i,j^-) \leq F_1(i)H(i^-,j^-) \\ &\iff LTD(X_2|X_1) \end{aligned}$$

et

$$\begin{aligned} A_{ij} + B_{ij} \leq 0 &\iff F_1(i^-)H(i,j) \leq F_1(i)H(i^-,j) \\ &\iff LTD(X_2|X_1). \end{aligned}$$

La preuve de la deuxième équivalence s'effectue de façon similaire. ◇

Le corollaire suivant, conséquence directe de la proposition précédente, généralise la Proposition 3.3 de Genest et Capéraà (1993) [6] au cas discret.

Corollaire 3.1. *Soient X_1 et X_2, des variables aléatoires discrètes à valeurs dans $\mathbb{Z}$ et de loi jointe H. On a alors que $RTI(X_2|X_1)$ et $LTD(X_2|X_1)$ impliquent que*

$$\rho_H \geq \tau_H \geq 0.$$

En suivant le même raisonnement, on parvient à montrer que

$$LTI(X_2|X_1) \iff LTI(X_2^\star|X_1^\star)$$

et

$$RTD(X_2|X_1) \Longleftrightarrow RTD(X_2^\star|X_1^\star).$$

Ainsi, $RTD(X_2|X_1)$ et $LTI(X_2|X_1)$ impliquent que

$$\rho_H \leq \tau_H \leq 0.$$

3.6 Construction de la copule de $(X_1^\star, X_2^\star)$

Soient X_1 et X_2, des variables aléatoires discrètes à valeurs dans $\mathbb{Z}$. Le but ici est d'obtenir l'unique copule associée aux variables continues $X_1^\star = X_1 + U_1$ et $X_2^\star = X_2 + U_2$. D'abord, dans le cas où les variables aléatoires U_1 et U_2 sont indépendantes, on a

$$H^\star(x_1, x_2) = \sum_{i=-\infty}^{+\infty} \sum_{j=-\infty}^{+\infty} (x_1 - i)(x_2 - j)h(i,j).$$

D'après le théorème de Sklar, la copule $C^\star$ associée à $H^\star$ est

$$C^\star(u_1, u_2) = H^\star\left\{(F_1^\star)^{-1}(u_1), (F_2^\star)^{-1}(u_2)\right\}, \tag{3.6}$$

où $F_1^\star(x) = H^\star(x, \infty)$ et $F_2^\star(x) = H^\star(\infty, x)$ sont les fonctions de répartition marginales. On peut montrer que $F_i^\star(x) = (x - i)f_i(j) + F_i(j)$ pour $x \in [j-1, j]$, $i = 1, 2$. Ceci implique que

$$F_i^\star(u) = j + \frac{u - F_i(j)}{f_i(j)}, \quad u \in [F_i(j-1), F_i(j)], \quad i = 1, 2.$$

En injectant ces deux dernières égalités dans l'équation (3.6), on obtient

$$\begin{aligned} C^{\star}(u_1,u_2) &= \frac{\{u_1-F_1(i^-)\}\{u_2-F_2(j^-)\}h(i,j)}{f_1(i)f_2(j)} \\ &+\frac{\{u_1-F_1(i^-)\}\{H(i,j^-)-H(i^-,j^-)\}}{f_1(i)} \\ &+\{u_2-F_2(j^-)\}\{H(i,j^-)-H(i^-,j^-)\}f_2(j) \\ &+H(i^-,j^-), \end{aligned}$$

où $(u_1,u_2)\in[F_1(i^-),F_1(i)]\times[F_2(j^-),F_2(j)]$.

3.7 La loi de Poisson bivariée

Plusieurs modèles ont été proposés pour modéliser la dépendance entre différents types de sinistres. Par exemple, en assurance non vie, on veut parfois expliquer le nombre de sinistres survenues sur une période donnée. Les observations étant de nature discrète, des modèles de dépendance pour des variables à valeurs dans des ensembles dénombrables doivent être proposés.

Le modèle de Poisson bivarié a été utilisé, sous différentes versions, par Amabagaspitiya (1998), Wang (1998) et Cossette et Marceau (2000) [11]. La particularité de ces lois est que les distributions des variables individuelles sont dans la famille bien connue des lois de Poisson.

Le but de ce chapitre est d'étudier les mesures de dépendance de Kendall et de Spearman pour le modèle de Poisson à deux variables avec chocs communs. Pour être précis, soient Y_1, Y_2 et Z, des variables aléatoires indépendantes distribuées selon la loi de Poisson de paramètres respectifs λ_1, λ_2 et α. Le

modèle de Poisson proposé représente la loi jointe du couple (X_1, X_2), où

$$X_1 = Y_1 + Z \quad \text{et} \quad X_2 = Y_2 + Z.$$

Explicitement, on a

$$\begin{aligned} H_{\alpha,\lambda_1,\lambda_2}(i,j) &= \mathrm{P}(X_1 \leq i, X_2 \leq j) \\ &= \mathrm{P}(Y_1 + Z \leq i, Y_2 + Z \leq j) \\ &= \mathrm{P}(Y_1 \leq i - Z, Y_2 \leq j - Z) \\ &= \sum_{z=0}^{\infty} \mathrm{P}(Y_1 \leq i - z, Y_2 \leq j - z | Z = z)\mathrm{P}(Z = z). \end{aligned}$$

Finalement, comme Y_1, Y_2 et Z sont indépendantes, on déduit

$$\begin{aligned} H_{\alpha,\lambda_1,\lambda_2}(i,j) &= \sum_{z=0}^{\infty} \mathrm{P}(Y_1 \leq i - z)\mathrm{P}(Y_2 \leq j - z)\mathrm{P}(Z = z) \\ &= \sum_{z=0}^{\infty} F_{\lambda_1}(i - z)F_{\lambda_2}(j - z)f_{\alpha}(z), \end{aligned}$$

où $f_{\alpha}(z) = \mathrm{P}(Z = z) = \alpha^z e^{-\alpha}/z!$ et F_{λ_i} est la fonction de répartition de Y_i, $i = 1, 2$.

Il serait intéressant d'étudier l'effet des paramètres λ_1, λ_2 et α sur la dépendance entre les variables dans ce modèle de Poisson bivarié. Spécifiquement, on étudiera la relation entre le paramètre α et les mesures de dépendance τ et ρ. Pour cela, on pose

$$m_1 = \lambda_1 + \alpha \quad \text{et} \quad m_2 = \lambda_2 + \alpha.$$

Ainsi, la fonction de répartition jointe du vecteur aléatoire (X_1, X_2) s'écrit

$$H_{\alpha}(i,j) = \sum_{z=0}^{\infty} F_{m_1-\alpha}(i - z)F_{m_2-\alpha}(j - z)f_{\alpha}(z). \tag{3.7}$$

Ce changement de notation entraîne que H_α est un modèle de Poisson dont les marges sont des lois de Poisson de paramètres respectifs m_1 et m_2. Par conséquent, le paramètre α est un indice de dépendance. Le résultat suivant indique que plus α prend des valeurs élevées, plus la dépendance positive est forte. Ceci a des conséquences directes sur le comportement du tau de Kendall et du rho de Spearman.

Proposition 3.7. *Soient H_{α_1} et H_{α_2}, des distributions de Poisson décrites par l'équation* (3.7). *Alors*

$$\alpha_1 \leq \alpha_2 \Rightarrow H_{\alpha_1}(i,j) \leq H_{\alpha_2}(i,j), \quad \textit{pour tout } (i,j) \in \mathbb{N} \times \mathbb{N}.$$

Par conséquent, si τ_{α_i} et ρ_{α_i} sont le tau de Kendall et le rho de Spearman associés à H_{α_i}, $i = 1, 2$, on a

$$\alpha_1 \leq \alpha_2 \Rightarrow \tau_{\alpha_1} \leq \tau_{\alpha_2} \quad \textit{et} \quad \alpha_1 \leq \alpha_2 \Rightarrow \rho_{\alpha_1} \leq \rho_{\alpha_2}.$$

Démonstration. Pour montrer le résultat, il suffit de montrer que H_α est croissante en fonction de α. En appliquant la règle de dérivation en chaîne, on obtient

$$\begin{aligned}
\frac{\partial}{\partial\alpha} H_\alpha(i,j) &= \sum_{z=0}^{\infty} \dot{F}_{m_1-\alpha}(i-z) F_{m_2-\alpha}(j-z) f_\alpha(z) \\
&\quad + \sum_{z=0}^{\infty} F_{m_1-\alpha}(i-z) \dot{F}_{m_2-\alpha}(j-z) f_\alpha(z) \\
&\quad + \sum_{z=0}^{\infty} F_{m_1-\alpha}(i-z) F_{m_2-\alpha}(j-z) \left\{ f_\alpha(z-1) - f_\alpha(z) \right\}
\end{aligned}$$

où

$$\dot{F}_{m_i-\alpha} = \frac{\partial}{\partial\alpha} F_{m_i-\alpha}, \quad i = 1, 2.$$

Pour simplifier ce qui suit, on définit $a_z(x) = F_{m_1-\alpha}(x-z)$ et $b_z(x) = F_{m_2-\alpha}(x-z)$. On montre facilement, par quelques manipulations algébriques simples, que

$$\dot{F}_{m_1-\alpha}(x-z) = a_z(x) - a_{z+1}(x) \quad \text{et} \quad \dot{F}_{m_2-\alpha}(x-z) = b_z(x) - b_{z+1}(x).$$

On a donc, après quelques calculs,

$$\begin{aligned}\frac{\partial}{\partial\alpha}H_\alpha(i,j) &= \sum_{z=0}^{\infty}\{a_z(i)b_z(j) - a_{z+1}(i)b_z(j) - a_z(i)b_{z+1}(j)\}\, f_\alpha(z) \\ &\quad + \sum_{z=1}^{\infty} a_z(i)b_z(j)f_\alpha(z-1).\end{aligned}$$

Comme

$$\sum_{z=1}^{\infty} a_z(i)b_z(j)f_\alpha(z-1) = \sum_{z=0}^{\infty} a_{z+1}(i)b_{z+1}(j)f_\alpha(z),$$

on conclut que

$$\begin{aligned}\frac{\partial}{\partial\alpha}H_\alpha(i,j) &= \sum_{z=0}^{\infty}\{a_z(i)b_z(j) - a_{z+1}(i)b_z(j) - a_z(i)b_{z+1}(j) + a_{z+1}(i)b_{z+1}(j)\}\, f_\alpha(z) \\ &= \sum_{z=0}^{\infty}\{a_z(i) - a_{z+1}(i)\}\{b_z(j) - b_{z+1}(j)\}\, f_\alpha(z) \\ &\geq 0\end{aligned}$$

car $a_z(i) - a_{z+1}(i) \geq 0$ et $b_z(i) - b_{z+1}(i) \geq 0$ pour tout $i \in \mathbb{N}$. Ceci montre que H_α est croissante en fonction de α. Finalement, les conclusions sur τ_α et ρ_α sont immédiates, invoquant le caractère monotone de ces mesures d'association. ◇

Dans la suite, nous illustrons les résultats de monotonicité par figures qui représentent le comportement du tau de Kendall en fonction du paramètre de dépendance α.

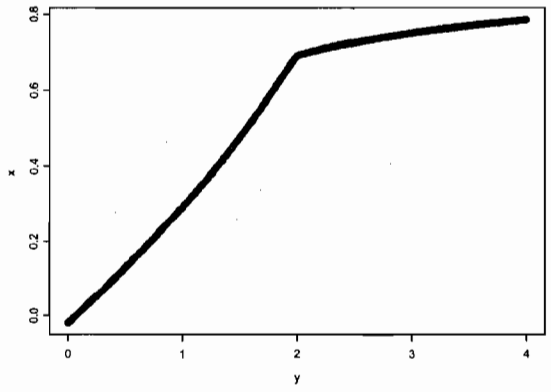

Figure 3.1: Comportement du tau de Kendall en fonction du paramètre α, $y = \alpha$ et $x = \tau$ et $m_1 = m_2 = 2$

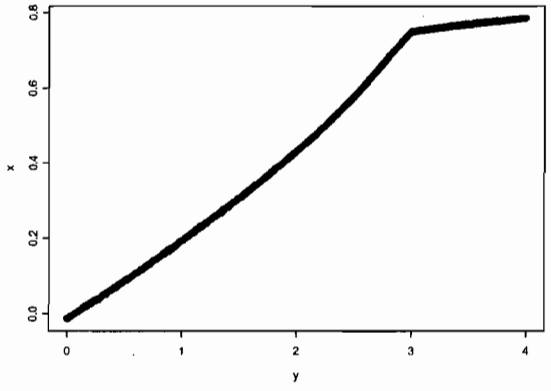

Figure 3.2: Comportement du tau de Kendall en fonction du paramètre α, $y = \alpha$ et $x = \tau$ et $m_1 = m_2 = 3$

CHAPITRE 4

EXTENSIONS MULTIVARIÉES DE TAU ET RHO

L'analyse des données multivariées discrètes est d'une grande importance dans une multitude de domaines. Citons à titre d'exemple le cas de l'assurance non-vie où l'étude de dépendance entre les différentes branches de réclamation est considérée essentielle.

Dans ce chapitre nous allons généraliser le tau de Kendall et le rho de Spearman discrets établis au chapitre 3 au cas multivariées. L'étude de la monotonicité de ses paramètres sera examinée. Afin de faciliter l'interprétation de ces nouvelles mesures, nous proposerons des transformations de sorte que les supports de ces paramètres coïncident avec l'intervalle $[0, 1]$. Nous terminons ce chapitre par des illustrations faisons intervenir certains modèles discrets. Un intérêt particulier sera porté aux modèles de chocs basés sur la loi de Poisson mulivariée.

4.1 Construction du tau de Kendall multivarié pour des données ordinales

Pour déduire une expression pour le tau de Kendall dans le cas continu multivarié, on avait imposé que cette mesure s'écrive comme une transformation affine de $\mathrm{E}_H\{H(X)\}$, où $X = (X_1, \ldots, X_d)$ est un vecteur aléatoires de loi H prenant ses valeurs dans $\mathbb{R}^d$. On en a alors déduit un tau de Kendall multivarié, à savoir

$$\tau^d(X) = \frac{2^d \mathrm{E}_H\{H(X)\} - 1}{2^{d-1} - 1}$$

Pour en arriver à cette expression, on a imposé que τ^d prenne les valeurs 0 et $+1$ respectivement pour l'indépendance et la dépendance positive parfaite.

Pour le cas de variables aléatoires définies sur une échelle ordinale, ce raisonnement n'est plus valide car la valeur du tau de Kendall lors d'une dépendance positive parfaite peut ne pas atteindre la valeur $+1$. Une autre approche sera alors employée dans les développements subséquents. D'abord, rappelons que dans le cas $d = 2$, on peut écrire le tau de Kendall par l'expression

$$\begin{aligned} \tau &= \mathrm{E}\left\{H(X_1^-, X_2^-)\right\} + \mathrm{E}\left\{H(X_1^-, X_2)\right\} \\ &\quad + \mathrm{E}\left\{H(X_1, X_2^-)\right\} + \mathrm{E}\left\{H(X_1, X_2)\right\} - 1. \end{aligned}$$

D'une certaine façon, cette formule tient compte de toutes les combinaisons d'égalités possibles parmi deux vecteurs aléatoires.

Pour étendre cette expression au cas $d = 3$, prenons un vecteur aléatoire $X = (X_1, X_2, X_3)$ de fonction de répartition jointe H. On pourrait définir

un tau de Kendall en fonction de

$$\begin{aligned}\ell(H) &= \mathrm{E}_H\left\{H(X_1^-, X_2^-, X_3^-)\right\} + \mathrm{E}_H\left\{H(X_1^-, X_2^-, X_3)\right\} \\ &\quad + \mathrm{E}_H\left\{H(X_1^-, X_2, X_3^-)\right\} + \mathrm{E}_H\left\{H(X_1^-, X_2, X_3)\right\} \\ &\quad + \mathrm{E}_H\left\{H(X_1, X_2^-, X_3^-)\right\} + \mathrm{E}_H\left\{H(X_1, X_2^-, X_3)\right\} \\ &\quad + \mathrm{E}_H\left\{H(X_1, X_2, X_3^-)\right\} + \mathrm{E}_H\left\{H(X_1, X_2, X_3)\right\}.\end{aligned}$$

Plus généralement, pour le cas à d variables, soit l'ensemble $\mathcal{E} = \{<, \leq\}^d$. Les éléments de $\mathcal{E}$ sont ainsi des vecteurs à d dimensions dont chaque composante est soit $<$ ou $\leq$. Comme exemples d'éléments de $\mathcal{E}$, on a

$$\tilde{<} = (<, \leq, \ldots, <) \quad \text{ou} \quad \tilde{<} = (\leq, \leq, \ldots, \leq).$$

Cette notation servira à définir le tau de Kendall ainsi que le rho de Spearman discrets multivariés. Quand $d = 2$, on a $\mathcal{E} = \{(<, <), (<, \leq), (\leq, <), (\leq, \leq)\}$. En outre, on notera $H_{\tilde{<}}(x) = \mathrm{P}(X \tilde{<} x)$, où $\tilde{<} \in \mathcal{E}$. Pour poursuivre l'idée précédente, on propose un tau de Kendall multivarié adapté pour des données ordinale qui sera fonction de

$$\ell(H) = \sum_{\tilde{<} \in \mathcal{E}} \mathrm{E}\left\{H_{\tilde{<}}(X)\right\}.$$

À noter que dans le cas continu, $\ell(H) = 2^d \mathrm{E}\{H(X)\}$. Ainsi, afin de retrouver comme cas particulier l'expression du tau de Kendall pour des variables continues, on définit

$$\tau^d = \frac{1}{2^{d-1} - 1} \sum_{\tilde{<} \in \mathcal{E}} \mathrm{E}\left\{H_{\tilde{<}}(X)\right\} - \frac{1}{2^{d-1} - 1}. \tag{4.1}$$

Exemple 4.1. *Pour illustrer le calcul de cette version discrète du tau de Kendall, soit un vecteur aléatoire* (X_1, X_2, X_3) *tel que*

$$P_{ijk} = \frac{1}{8} + (-1)^{i+j+k} A, \quad i, j, k \in \{0, 1\},$$

où $\mathrm{P}_{ijk} = \mathrm{P}(X_1 = i, X_2 = j, X_3 = k)$ *et* $0 \leq A \leq 1/8$. *De cette façon, chacune des marges est une loi Bernoulli de probabilité de succès* $1/2$. *On a donc*

$$\begin{array}{llll} \mathrm{P}_{000} = 1/8 + A & \mathrm{P}_{001} = 1/8 - A & \mathrm{P}_{010} = 1/8 - A & \mathrm{P}_{011} = 1/8 + A \\ \mathrm{P}_{101} = 1/8 + A & \mathrm{P}_{100} = 1/8 - A & \mathrm{P}_{110} = 1/8 + A & \mathrm{P}_{111} = 1/8 - A. \end{array}$$

Afin d'appliquer la formule (4.1), *notons d'abord que*

$$\begin{aligned} \mathrm{E}_H \{H(X_1, X_2, X_3)\} &= \sum_{(i,j,k) \in \{0,1\}^3} H(i, j, k) \mathrm{P}_{ijk} \\ &= H(0,0,0)\mathrm{P}_{000} + H(0,0,1)\mathrm{P}_{001} + H(1,0,0)\mathrm{P}_{100} \\ &\quad + H(0,1,0)\mathrm{P}_{010} + H(1,1,0)\mathrm{P}_{110} + H(1,0,1)\mathrm{P}_{101} \\ &\quad + H(0,1,1)\mathrm{P}_{011} + H(1,1,1)\mathrm{P}_{111} \\ &= 27/64 + A^2. \end{aligned}$$

Dans la même veine,

$$\begin{aligned}
\mathrm{E}_H\left\{H(X_1^-, X_2, X_3)\right\} &= \sum_{(i,j,k)\in\{0,1\}^3} H(i^-, j, k)\mathrm{P}_{ijk} \\
&= H(0,0,0)\mathrm{P}_{100} + H(0,0,1)\mathrm{P}_{101} \\
&\quad + H(0,1,0)\mathrm{P}_{110} + H(0,1,1)\mathrm{P}_{111} \\
&= \mathrm{P}_{000}\mathrm{P}_{100} + (\mathrm{P}_{000} + \mathrm{P}_{001})\mathrm{P}_{101} + (\mathrm{P}_{000} + \mathrm{P}_{010})\mathrm{P}_{110} \\
&\quad + (\mathrm{P}_{000} + \mathrm{P}_{001} + \mathrm{P}_{010} + \mathrm{P}_{011})\mathrm{P}_{111} \\
&= \left(\frac{1}{64} - A^2\right) + \left(\frac{1}{32} + \frac{A}{4}\right) + \left(\frac{1}{32} + \frac{A}{4}\right) + \left(\frac{1}{16} + \frac{A}{2}\right) \\
&= \frac{9}{64} + A - A^2.
\end{aligned}$$

Les mêmes types de calculs amènent

$$\mathrm{E}_H\left\{H(X_1, X_2^-, X_3)\right\} = \mathrm{E}_H\left\{H(X_1, X_2, X_3^-)\right\} = \frac{9}{64} + A - A^2.$$

Aussi,

$$\begin{aligned}
\mathrm{E}_H\left\{H(X_1^-, X_2^-, X_3)\right\} &= \sum_{(i,j,k)\in\{0,1\}^3} H(i^-, j^-, k)\mathrm{P}_{ijk} \\
&= H(0,0,0)\mathrm{P}_{110} + H(0,0,1)\mathrm{P}_{111} \\
&= \mathrm{P}_{000}\mathrm{P}_{110} + (\mathrm{P}_{000} + \mathrm{P}_{001})\mathrm{P}_{111} \\
&= \left(\frac{1}{8} + A\right)^2 + \left(\frac{1}{32} - \frac{A}{4}\right) \\
&= \frac{3}{64} + A^2,
\end{aligned}$$

et

$$\mathrm{E}_H\left\{H(X_1^-, X_2, X_3^-)\right\} = \mathrm{E}_H\left\{H(X_1, X_2^-, X_3^-)\right\} = \frac{3}{64} + A^2.$$

Enfin,

$$\begin{aligned} \mathrm{E}_H\left\{H(X_1^-, X_2^-, X_3^-)\right\} &= \sum_{(i,j,k)\in\{0,1\}^3} H(i^-, j^-, k^-)\mathrm{P}_{ijk} \\ &= H(0,0,0)\mathrm{P}_{111} \\ &= \frac{1}{64} - A^2. \end{aligned}$$

En faisant le total, on trouve

$$\ell(H) = 1 + 6A - 6A^2,$$

d'où

$$\tau^3 = \frac{1}{3}\left(1 + 6A - 6A^2\right) - \frac{1}{3} = 2A(1 - A).$$

Lorsque $A = 0$, la probabilité dans chaque cellule est $1/8 = 1/2 \times 1/2 \times 1/2$, c'est-à-dire le produit des marges. Comme cette situation caractérise l'indépendance, c'est sans surprise qu'on obtient $\tau^3 = 0$. Aussi, puisque $2A(1-A)$ est une fonction croissante sur l'intervalle $[0, 1/8]$, la valeur maximale du tau de Kendall pour ce modèle est $\tau^3 = 7/32$.

4.2 Construction du rho de Spearman multivarié pour des données ordinales

Un raisonnement semblable à celui suivi pour établir le tau de Kendall discret servira à élaborer un rho de Spearman adapté au cas discret. D'abord, à noter

que dans le cas bivarié discret,

$$\begin{aligned}\rho &= \mathrm{E}_\Pi\left\{H(X_1^-, X_2^-)\right\} + \mathrm{E}_\Pi\left\{H(X_1^-, X_2)\right\} \\ &\quad + \mathrm{E}_\Pi\left\{H(X_1, X_2^-)\right\} + \mathrm{E}_\Pi\left\{H(X_1, X_2)\right\} - 3.\end{aligned}$$

Ainsi, se basant sur la version multivariée de cette mesure, on propose de bâtir une mesure à partir de

$$\ell_2(H) = \sum_{\underset{\sim}{\preceq} \in \mathcal{E}} \mathrm{E}_\Pi\left\{H_{\underset{\sim}{\preceq}}(X)\right\}.$$

Spécifiquement, on définit

$$\rho^d = (d+1)\left[\sum_{\underset{\sim}{\preceq} \in \mathcal{E}} \mathrm{E}_\Pi\left\{H_{\underset{\sim}{\preceq}}(X)\right\} - 1\right] \Big/ \left\{2^d - (d+1)\right\}, \tag{4.2}$$

où $X = (X_1, \ldots, X_d)$ est un vecteur de loi H. Cette formule généralise le rho de Spearman multivarié proposé précédemment pour le cas de variables continues. À remarquer que $\rho^d = 0$ dans le cas où les variables sont indépendantes.

Exemple 4.2. *Reprenons les données de l'exemple 4.1. D'abord, la formule pour $d = 3$ est*

$$\begin{aligned}\rho^3 &= \frac{3}{4}\left[\mathrm{E}_\Pi\left\{H(X_1, X_2, X_3)\right\} + \mathrm{E}_\Pi\left\{H(X_1^-, X_2, X_3)\right\}\right. \\ &\quad + \mathrm{E}_\Pi\left\{H(X_1, X_2^-, X_3)\right\} + \mathrm{E}_\Pi\left\{H(X_1, X_{,2}, X_3^-)\right\} \\ &\quad + \mathrm{E}_\Pi\left\{H(X_1^-, X_2^-, X_3)\right\} + \mathrm{E}_\Pi\left\{H(X_1^-, X_2, X_3^-)\right\} \\ &\quad \left. + \mathrm{E}_\Pi\left\{H(X_1, X_2^-, X_3^-)\right\} + \mathrm{E}_\Pi\left\{H(X_1^-, X_2^-, X_3^-)\right\} - 1\right].\end{aligned}$$

D'abord, à noter que les probabilités de chaque cellule sous l'indépendance est

$1/2 \times 1/2 \times 1/2 = 1/8$. *Ainsi,*

$$\begin{aligned}
8\mathrm{E}_\Pi\{H(X_1,X_2,X_3)\} &= \sum_{(i,j,k)\in\{0,1\}^3} H(i,j,k) \\
&= \mathrm{P}_{000} + (\mathrm{P}_{000}+\mathrm{P}_{001}) + (\mathrm{P}_{000}+\mathrm{P}_{010}) + (\mathrm{P}_{000}+\mathrm{P}_{100}) \\
&\quad + (\mathrm{P}_{000}+\mathrm{P}_{001}+\mathrm{P}_{010}+\mathrm{P}_{011}) + (\mathrm{P}_{000}+\mathrm{P}_{100}+\mathrm{P}_{001}+\mathrm{P}_{101}) \\
&\quad + (\mathrm{P}_{000}+\mathrm{P}_{100}+\mathrm{P}_{010}+\mathrm{P}_{110}) + 1 \\
&= 8\mathrm{P}_{000} + 4\mathrm{P}_{001} + 4\mathrm{P}_{010} + 4\mathrm{P}_{100} + 2\mathrm{P}_{011} + 2\mathrm{P}_{101} + 2\mathrm{P}_{110} + \mathrm{P}_{111} \\
&= \frac{27}{8} + A.
\end{aligned}$$

Également,

$$\begin{aligned}
8\mathrm{E}_\Pi\{H(X_1^-,X_2,X_3)\} &= \sum_{(i,j,k)\in\{0,1\}^3} H(i,j,k) \\
&= H(0,0,0) + H(0,0,1) + H(0,1,0) + H(0,1,1) \\
&= 4\mathrm{P}_{000} + 2\mathrm{P}_{001} + 2\mathrm{P}_{010} + \mathrm{P}_{011} \\
&= \frac{9}{8} + A.
\end{aligned}$$

Il s'ensuit directement que

$$8\mathrm{E}_\Pi\{H(X_1,X_2^-,X_3)\} = 8\mathrm{E}_\Pi\{H(X_1,X_2,X_3^-)\} = \frac{9}{8} + A.$$

Ensuite,

$$8\mathrm{E}_\Pi\{H(X_1^-,X_2^-,X_3)\} = H(0,0,0) + H(0,0,1) = 2\mathrm{P}_{000} + \mathrm{P}_{001} = \frac{3}{8} + A,$$

ce qui fait que

$$8\mathrm{E}_\Pi\{H(X_1^-,X_2,X_3^-)\} = 8\mathrm{E}_\Pi\{H(X_1,X_2^-,X_3^-)\} = \frac{3}{8} + A.$$

Enfin,

$$8\mathrm{E}_\Pi\{H(X_1^-,X_2^-,X_3^-)\} = H(0,0,0) = \frac{1}{8} + A.$$

En rassemblant le tout, on trouve d'abord $\ell_2(H) = 1 + A$. *Finalement, ceci permet de calculer*

$$\rho^3 = \frac{3A}{4}.$$

Sous l'indépendance, c'est-à-dire quand $A = 0$, *on a* $\rho^3 = 0$, *tel qu'attendu. La valeur maximale du rho de Spearman pour ce modèle est* $\rho^3 = 3/32$.

4.3 Monotonocité du tau de Kendall et du rho de Spearman multivariées

De la même façon que cela a été fait au Chapitre 2, on montrera ici la propriété de monotonicité des versions multivariées discrètes du tau de Kendall et du rho de Spearman. Pour ce faire, soient deux vecteurs aléatoires $X = (X_1, \ldots, X_d)$ et $Y = (Y_1, \ldots, Y_d)$ de fonctions de répartitions jointes respectives H et G. Les implications

$$X \prec_c Y \Longrightarrow \tau_H \leq \tau_G \quad \text{et} \quad X \prec_c Y \Longrightarrow \rho_H \leq \rho_G$$

seront démontrées dans la suite.

Le lemme suivant sera instrumental dans les preuves subséquentes.

Lemme 4.2. *Pour toute paire de fonctions de répartition d-variées* H *et* G *et n'importe quel* $\tilde{\prec} \in \mathcal{E}$, *on a*

$$\mathrm{E}_H \left\{ G_{\tilde{\prec}}(X) \right\} = \mathrm{E}_G \left\{ \bar{H}_{\tilde{\prec}}(X) \right\}.$$

Démonstration. Sans perte de généralité et afin de faciliter la lecture, supposons que

$$\tilde{\preceq} = (\underbrace{\leq, \ldots, \leq}_{k fois}, \underbrace{\leq, \ldots, \leq}_{d-k fois}).$$

Le résultat pour les autres vecteurs de $\mathcal{E}$ s'obtiendra de façon similaire.

D'abord,

$$\mathrm{E}_H \{G_{\tilde{\preceq}}(X)\} = \sum_{i \in \mathbb{Z}^d} G_{\tilde{\preceq}}(i) h(i).$$

Comme

$$G_{\tilde{\preceq}}(i) = \sum_{\ell \tilde{\preceq} i} g(\ell),$$

où g et h désignent les densités de probabilité associées respectivement aux fonctions de répartition G et H.

On déduit en utilisant le lemme de Fubini que

$$\begin{aligned} \mathrm{E}_H \{G_{\tilde{\preceq}}(X)\} &= \sum_{i \in \mathbb{Z}^d} \sum_{\ell \tilde{\preceq} i} g(\ell) h(i) \\ &= \sum_{\ell \in \mathbb{Z}^d} \left\{ \sum_{i \tilde{\succeq} \ell} h(i) \right\} g(\ell) \\ &= \sum_{\ell \in \mathbb{Z}^d} \bar{H}_{\tilde{\preceq}}(\ell) g(\ell) \\ &= \mathrm{E}_G \left\{ \bar{H}_{\tilde{\preceq}}(X) \right\}. \end{aligned}$$

Proposition 4.1. *Soient $X = (X_1, \ldots, X_d)$ de fonction de répartition H et $Y = (Y_1, \ldots, Y_d)$ de fonction de répartition G. Alors*

$$X \prec_c Y \Longrightarrow \tau_H^d \leq \tau_G^d. \tag{4.3}$$

Démonstration. Prenons $X \leqslant_c Y$, c'est-à-dire que pour tout x et n'importe quel $\tilde{\prec} \in \mathcal{E}$, $H_{\tilde{\prec}}(x) \leq G_{\tilde{\prec}}(x)$. Il est alors clair que

$$\mathrm{E}_H \{H_{\tilde{\prec}}(X)\} \leq \mathrm{E}_H \{G_{\tilde{\prec}}(X)\}.$$

Ensuite, du lemme (4.2), $\mathrm{E}_H \{G_{\tilde{\prec}}(X)\} = \mathrm{E}_G \{\bar{H}_{\tilde{\prec}}(X)\}$. Maintenant, la définition de l'ordre de concordance implique que pour tout x et n'importe quel $\tilde{\prec} \in \mathcal{E}$, $\bar{H}_{\tilde{\prec}}(x) \leq \bar{G}_{\tilde{\prec}}(x)$. Donc, $\mathrm{E}_G \{\bar{H}_{\tilde{\prec}}(X)\} \leq \mathrm{E}_G \{\bar{G}_{\tilde{\prec}}(X)\}$. Il s'ensuit que

$$\mathrm{E}_H \{H_{\tilde{\prec}}(X)\} \leq \mathrm{E}_G \{\bar{G}_{\tilde{\prec}}(X)\} = \mathrm{E}_G \{G_{\tilde{\prec}}(X)\},$$

où la dernière égalité provient d'une autre application du lemme 4.2. La conclusion annoncée pour le tau de Kendall est immédiate. ◇

La proposition suivante amène un résultat semblable pour le rho de Spearman multivarié discret.

Proposition 4.2. *Soient $X = (X_1, \ldots, X_d)$ de fonction de répartition H et $Y = (Y_1, \ldots, Y_d)$ de fonction de répartition G. Alors*

$$X \prec_c Y \Longrightarrow \rho_H^d \leq \rho_G^d. \tag{4.4}$$

Démonstration. Comme $H_{\tilde{\prec}}(x) \leq G_{\tilde{\prec}}(x)$ pour tout x et n'importe quel $\tilde{\prec} \in \mathcal{E}$, on a $\mathrm{E}_\Pi \{H_{\tilde{\prec}}(X)\} \leq \mathrm{E}_\Pi \{G_{\tilde{\prec}}(X)\}$, d'où on déduit facilement, de par la définition même du rho de Spearman multivarié discret, que $\rho_H^d \leq \rho_G^d$. ◇

À la lumière du chapitre 3, on va considérer le modèle de Poisson pour 3 variables avec chocs communs, le graphe ci-dessous nous montre le comportement du tau de Kendall en fonction du paramètre de dépendance α.

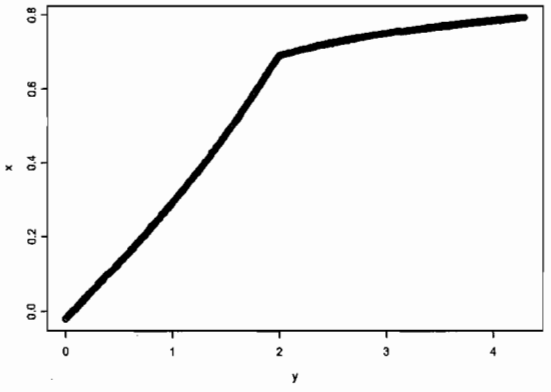

Figure 4.1: Comportement du tau de Kendall en fonction du paramètre α, $y = \alpha$ et $x = \tau$ et $m_1 = m_2 = m_3 = 2$

On remarque que le tau de Kendall est croissant, ce résultat n'a pas été démontré mathématiquement.

4.4 Versions corrigées

L'interprétation des valeurs du tau de Kendall ou du rho de Spearman est difficile du fait que ces mesures n'atteignent pas nécessairement $+1$ en présence de dépendance positive parfaite, c'est-à-dire quand la copule sous-jacente à une loi H est la borne supérieure de Fréchet $M^d(u_1, \ldots, u_d) = \min(u_1, \ldots, u_d)$. Dans cette section, on cherchera à corriger cette situation.

Soit une loi H de marges $F_1, \ldots, F_d$. On notera par $\tau^d_{\max}(F_1, \ldots, F_d)$ et

$\rho^d_{\max}(F_1, \ldots, F_d)$ le tau de Kendall et le rho de Spearman associés à la loi

$$G(x_1, \ldots, x_d) = M^d \{F_1(x_1), \ldots, F_d(x_d)\}.$$

La monotonicité du tau de Kendall et du rho de Spearman permet d'établir immédiatement que

$$\tau^d_H \leq \tau^d_{\max}(F_1, \ldots, F_d) \quad \text{et} \quad \rho^d_H \leq \rho^d_{\max}(F_1, \ldots, F_d).$$

Une façon de rendre le tau de Kendall et le rho de Spearman plus facile à interpréter consiste à utiliser les versions corrigées

$$\tilde{\tau}^d_H = \frac{\tau_H}{\tau_{\max}(F_1, \ldots, F_d)} \quad \text{et} \quad \tilde{\rho}^d_H = \frac{\rho_H}{\rho^d_{\max}(F_1, \ldots, F_d)}.$$

Ces nouvelles mesures atteignent ainsi $+1$ lors de la dépendance positive parfaite.

Exemple 4.3. *Lorsque le modèle de dépendance sous-jacent est la copule de Fréchet M^3 et que les marges sont Bernoulli de probabilité de succès $1/2$, alors la masse de probabilité est concentrée sur $\mathrm{P}_{000} = 1/2$ et $\mathrm{P}_{111} = 1/2$. Premièrement,*

$$\begin{aligned} \mathrm{E}_{M^3}\left\{M^3(X_1, X_2, X_3)\right\} &= M^3(0,0,0)\mathrm{P}_{000} + M^3(1,1,1)\mathrm{P}_{111} \\ &= (1/2)(1/2) + (1/2) \\ &= 3/4. \end{aligned}$$

Ensuite,

$$\mathrm{E}_{M^3}\left\{M^3(X_1^-, X_2, X_3)\right\} = M^3(0,1,1)\mathrm{P}_{111} = (1/2)(1/2) = 1/4.$$

De la même façon,

$$\mathrm{E}_{M^3}\left\{M^3(X_1, X_2^-, X_3)\right\} = \mathrm{E}_{M^3}\left\{M^3(X_1, X_2, X_3^-)\right\} = 1/4.$$

Aussi,

$$\mathrm{E}_{M^3}\left\{M^3(X_1^-, X_2^-, X_3)\right\} = M^3(0,0,1)\mathrm{P}_{111} = (1/2)(1/2) = 1/4,$$

ce qui fait que

$$\mathrm{E}_{M^3}\left\{M^3(X_1^-, X_2, X_3^-)\right\} = \mathrm{E}_{M^3}\left\{M^3(X_1, X_2^-, X_3^-)\right\} = 1/4.$$

Enfin,

$$\mathrm{E}_{M^3}\left\{M^3(X_1^-, X_2^-, X_3^-)\right\} = M^3(0,0,0)\mathrm{P}_{111} = (1/2)(1/2) = 1/4,$$

En rassemblant le tout, on trouve

$$\tau^3_{\max}(F_1, F_2, F_3) = \frac{1}{3}\left(\frac{10}{4} - 1\right) = \frac{1}{2}.$$

Il est donc impossible dans un modèle de marges Bernoulli identiques que le tau de Kendall surpasse la valeur $1/2$. *En particulier, pour le modèle de l'exemple 4.1, la version corrigée est donnée par*

$$\tilde{\tau}^3 = \frac{2A(1-A)}{1/2} = 4A(1-A).$$

À noter toutefois que comme $A \leq 1/8$, *on a* $\tilde{\tau}^3 \leq 7/16$. *Ceci est dû au fait que ce modèle ne permet pas d'atteindre la dépendance positive parfaite.*

CHAPITRE 5

UN TAU DE KENDALL MULTIVARIÉ EMPIRIQUE POUR DES DONNÉES ORDINALES

5.1 Définition d'une version empirique du tau de Kendall théorique discret multivarié

Pour un vecteur $X = (X_1, \ldots, X_d)$ dont les composantes sont des variables aléatoires définies sur un support discret et dont la fonction de répartition jointe est H, on a défini

$$\tau_H^d = \frac{1}{2^{d-1}-1}\left[\sum_{\tilde{\prec}\in\mathcal{E}} \mathrm{E}\left\{H_{\tilde{\prec}}(X)\right\} - 1\right]$$

Pour estimer τ^d avec des observations, on trouvera des estimateurs pour chacun des termes de la sommation. D'abord, notons que pour tout $\tilde{\prec} \in \mathcal{E}$,

$$\mathrm{E}_H\left\{H_{\tilde{\prec}}(X)\right\} = \mathrm{P}_H\left(X \tilde{\prec} X'\right),$$

où $X' = (X_1', \ldots, X_d')$ est indépendant de X et distribué selon la loi H.

Soit maintenant $X_1, \ldots, X_n$, où $X_i = (X_{i1}, \ldots, X_{id})$, un échantillon de taille n. On estime $\mathrm{E}_H \{H_{\tilde{\preceq}}(X)\}$ sans biais par

$$\hat{E}\{H_{\tilde{\preceq}}(X)\} = \binom{n}{2}^{-1} \sum_{i<j} \{\mathbf{1}(X_i \tilde{\preceq} X_j) + \mathbf{1}(X_j \tilde{\preceq} X_i)\}.$$

Ainsi, l'estimation proposée pour τ_H^d est

$$\tau_n^d = \frac{1}{2^{d-1}-1} \left[\binom{n}{2}^{-1} \sum_{\tilde{\preceq} \in \mathcal{E}} \sum_{i<j} \{\mathbf{1}(X_i \tilde{\preceq} X_j) + \mathbf{1}(X_j \tilde{\preceq} X_i)\} - 1 \right]$$

Exemple 5.1. *Afin d'illustrer le calcul de τ_n^d, soit un échantillon $X_1, \ldots, X_6$ de taille $n = 6$, dans le tableau suivant, on présente les valeurs des composantes de l'échantillon qu'on a choisi. Remarquons que dans ce cas, on a*

Table 5.1: Illustration à des données aléatoires discrètes

observation i	X_{i1}	X_{i2}	X_{i3}
$i = 1$	2	3	5
$i = 2$	4	6	2
$i = 3$	1	0	5
$i = 4$	5	7	12
$i = 5$	5	8	12
$i = 6$	9	10	12

8 termes $\hat{\mathrm{E}}\left\{H_{\tilde{<}}(X)\right\}$ *à estimer, avec* $\tilde{<} \in \mathcal{E} = (<, \leq)^3$*, notons d'abord que*

$$\begin{aligned}\hat{\mathrm{E}}\left\{H_{(\leq,\leq,\leq)}(X)\right\} &= 1/45 \sum_{i<j<6} \mathbf{1}\left(X_{i1} \leq X_{j1}, X_{i2} \leq X_{j2}, X_{i3} \leq X_{j3}\right) \\ &+ \mathbf{1}\left(X_{j1} \leq X_{i1}, X_{j2} \leq X_{i2}, X_{j3} \leq X_{i3}\right) \\ &= 13/45\end{aligned}$$

Les mêmes types de calculs amènent

$$\hat{\mathrm{E}}\left\{H_{(<,\leq,\leq)(X)}\right\} = \hat{\mathrm{E}}\left\{H_{(<,<,\leq)}(X)\right\} = 12/45$$

Aussi

$$\hat{\mathrm{E}}\left\{H_{(\leq,\leq,<)}(X)\right\} = \hat{\mathrm{E}}\left\{H_{(<,\leq,<)}(X)\right\} = \hat{\mathrm{E}}\left\{H_{(\leq,<,<)}(X)\right\} = \hat{\mathrm{E}}\left\{H_{(<,<,<)}(X)\right\} = 9/45$$

Enfin,

$$\hat{\mathrm{E}}\left\{H_{(\leq,<,\leq)}(X)\right\} = 13/45$$

Ainsi, le tau de Kendall empirique est

$$\tau_6^3 = 41/45$$

5.2 Comportement asymptotique de la statistique τ_n^d

Le but de cette section est d'étudier le comportement asymptotique de τ_n^d. Une remarque importante est que cette statistique est un cas particulier de U-statistique. Cette classe de statistiques fut introduite par Hoeffding (1948) [12].

La définition générale d'une U-statistique est donnée dans ce qui suit.

Définition 5.8. *Soient $X_1, \ldots, X_n$, un échantillon tiré d'une population à valeurs dans $\mathbb{R}^d$. La statistique U_n est une U-statistique d'ordre 2 si on peut l'écrire sous la forme*

$$U_n = \binom{n}{2}^{-1} \sum_{i<j} \psi(X_i, X_j).$$

La fonction ψ est appelé le noyau de U_n.

Les résultats classiques concernant le Théorème de la limite centrale s'intéressent à l'étude de la somme de variables indépendantes et identiquement distribuées. Dans le cas d'une U-statistique, la sommation est effectuée sur des variables dépendantes, ce qui complique l'obtention de résultats asymptotiques. Cependant, par le méthode de projection, on peut montrer que celle-ci est équivalente asymptotiquement à une somme de variables aléatoires identiquement distribuées. Les résultats classiques sont dès lors applicables.

Théorème 5.2. *Si U_n est une U-statistique telle que $\mathrm{E}(U_n) = \theta$, alors $\sqrt{n}(U_n - \theta)$ est asymptotiquement normale de moyenne 0 et de variance $4\sigma_1^2$, où $\sigma_1^2 = \mathrm{var}\{\psi_1(X_1)\}$ et*

$$\psi_1(x) = \mathrm{E}_{X_2}\{\psi(x, X_2)\}.$$

La statistique τ_n^d est une U-statistique d'ordre 2. En effet,

$$\tau_n^d = \binom{n}{2}^{-1} \sum_{i<j} \psi(X_i, X_j),$$

où

$$\psi(X_i, X_j) = \left\{ \sum_{\tilde{z} \in \mathcal{E}} \{\mathbf{1}(X_i \tilde{<} X_j) + \mathbf{1}(X_j \tilde{<} X_i)\} - 1 \right\} \Big/ \left(2^{d-1} - 1\right).$$

On a donc

$$\begin{aligned}\psi_1(x_1) &= \mathrm{E}_{X_2}\left\{\psi(x_1, X_2)\right\} \\ &= \frac{1}{2^{d-1}-1}\left[\sum_{\tilde{<}\in\mathcal{E}}\left\{H_{\tilde{<}}(x_1) + \bar{H}_{\tilde{<}}(x_1)\right\} - 1\right].\end{aligned}$$

Ainsi,

$$4\sigma_1^2 = \left(\frac{2}{2^{d-1}-1}\right)^2 \mathrm{var}\left[\sum_{\tilde{<}\in\mathcal{E}}\left\{H_{\tilde{<}}(X_1) + \bar{H}_{\tilde{<}}(X_1)\right\}\right].$$

Donc, du Théorème 5.2, on conclut que $\sqrt{n}\left(\tau_n^d - \tau^d\right)$ converge vers une loi normale de moyenne 0 et de variance $4\sigma_1^2$.

Remarque 5.1. *Dans le cas continu et sous l'hypothèse d'indépendance, i.e.* $H(x_1, \ldots, x_d) = F_1(x_1)\cdots F_d(x_d)$, *on a*

$$\begin{aligned}4\sigma_1^2 &= \left(\frac{2}{2^{d-1}-1}\right)^2 \mathrm{var}\left[2^d\left\{H(X_1) + \bar{H}(X_1)\right\}\right] \\ &= \left(\frac{2^{d+1}}{2^{d-1}-1}\right)^2 \mathrm{var}\left\{\prod_{i=1}^d U_i + \prod_{i=1}^d (1-U_i)\right\} \\ &= \left(\frac{2^{d+1}}{2^{d-1}-1}\right)^2 \frac{1}{12^d}.\end{aligned}$$

Quand $d = 2$, *on retrouve le résultat bien connu qui stipule que la variance asymptotique du tau de Kendall bivarié est* $4/9$.

5.3 Estimation de la version corrigée du tau de Kendall

Définition 5.9. *Soit $X_1, X_2, \ldots$, une suite de variables aléatoires. On dit que la suite $\{X_n\}$ converge en probabilité vers μ si pour tout $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathrm{P}\left(|X_n - \mu| \geq \epsilon\right) = 0.$$

Définition 5.10. *Soit $X_1, X_2, \ldots$, une suite de variables aléatoires de fonction de répartition F_n. On dit que $\{X_n\}$ converge en loi vers X si*

$$\lim_{n\to\infty} F_n(x) = F(x)$$

pour tous les points où F est continue, avec $F(x) = \mathrm{P}\left(X \leq x\right)$.

Théorème 5.3. *(Slutsky) Soient $\{X_n\}$ et $\{Y_n\}$, deux suites de variables aléatoires. Si X_n converge en loi vers X et que Y_n converge en probabilité vers a, alors X_nY_n converge en loi vers aX.*

Théorème 5.4. *(Méthode Delta) Soit X_n, une suite de variables aléatoires telle que $\sqrt{n}\left(X_n - \mu\right)$ converge en loi vers $\mathcal{N}(0, \sigma^2)$. Si g est une fonction continue et dont la dérivée première existe, alors*

$$\sqrt{n}\left\{g(X_n) - g(\mu)\right\}$$

converge en loi vers une normale de moyenne 0 et de variance $\sigma^2\{g'(\mu)\}^2$.

La démonstration du résultat précédent s'obtient en effectuant un développement en série de Taylor d'ordre 1 et en appliquant le théorème de Slutsky.

Supposons maintenant que pour une collection donnée de marges $F_1, \ldots, F_d$, la valeur maximale du tau de Kendall dépende d'un paramètre ou d'un vecteur de paramètres θ. Autrement dit, $\tau^d_{\max}(F_1, \ldots, F_d) = \theta$. Si $\hat{\theta}$ est un estimateur de θ, alors

$$\tilde{\tau}^d_n = \frac{\tau^d_n}{\tau^d_{\max}(\hat{\theta})}$$

est un estimateur du tau de Kendall corrigé. On a

$$\begin{aligned}
\sqrt{n}\left(\tilde{\tau}^d_n - \tilde{\tau}^d\right) &= \sqrt{n}\left\{\frac{\tau^d_n}{\tau^d_{\max}(\hat{\theta})} - \frac{\tau^d}{\tau^d_{\max}(\theta)}\right\} \\
&= \frac{1}{\tau^d_{\max}(\hat{\theta})}\sqrt{n}\left(\tau^d_n - \tau^d\right) + \tau^d\sqrt{n}\left\{\frac{1}{\tau^d_{\max}(\hat{\theta})} - \frac{1}{\tau^d_{\max}(\theta)}\right\} \\
&= g(\hat{\theta})\sqrt{n}\left(\tau^d_n - \tau^d\right) + \tau^d\sqrt{n}\left\{g(\hat{\theta}) - g(\theta)\right\},
\end{aligned}$$

où $g(t) = \left\{\tau^d_{\max}(t)\right\}^{-1}$.

On a vu que $\sqrt{n}(\tau^d_n - \tau^d)$ converge en loi vers $\mathcal{N}(0, 4\sigma_1^2)$. Puisque $g(\hat{\theta})$ converge en probabilité vers $g(\theta)$, on a du Théorème de *Slutsky* que $g(\hat{\theta})\sqrt{n}(\tau^d_n - \tau^d)$ converge vers une loi normale de moyenne 0 et de variance $4\left\{g(\theta)\right\}^2\sigma_1^2$.

D'autre part, si on suppose que $\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, \sigma_2^2)$, alors en appliquant la méthode delta, on aura

$$\tau^d\sqrt{n}\left\{g(\hat{\theta}) - g(\theta)\right\}$$

converge vers une loi normale de moyenne 0 et de variance $(\tau^d)^2\sigma_2^2\{g'(\theta)\}^2$. En rassemblant le tout, on aura que $\sqrt{n}\left(\tilde{\tau}^d_n - \tilde{\tau}^d\right)$ converge vers une loi normale de moyenne 0 et de variance $\left(\sigma_2^2 + \left(\tau^d\right)^2\sigma_2^2\left\{g'\left(\theta\right)\right\}^2\right)$.
Ainsi, à partir de ces résultats, on peut effectuer des tests d'hypothèses sur la valeur du tau de Kendall corrigé (bilatéral et unilatéral), par la suite nous

posons $\Sigma^2 = \left(\sigma_2^2 + \left(\tau^d\right)^2 \sigma_2^2 \left\{g'\left(\theta\right)\right\}^2\right)$.
Un test bilatéral consiste à confronter les hypothèses

$$\mathcal{H}_0 : \tilde{\tau}^d = t \quad \text{vs} \quad \mathcal{H}_1 : \tilde{\tau}^d \neq t$$

On rejettera $\mathcal{H}_0$ lorsque

$$|\tau_n^d - t| > z_{\alpha/2}\frac{\Sigma}{\sqrt{n}},$$

alors qu'un test unilatéral à gauche consiste à confronter les hypothèses

$$\mathcal{H}_0 : \tilde{\tau}^d = t \quad \text{vs} \quad \mathcal{H}_1 : \tilde{\tau}^d < t.$$

on rejettera $\mathcal{H}_0$ lorsque

$$\tau_n^d < t + z_{\alpha}\frac{\Sigma}{\sqrt{n}},$$

tandis que pour un test unilatéral à droite, on rejettera $\mathcal{H}_0$ lorsque

$$\tau_n^d > t + z_{\alpha}\frac{\Sigma}{\sqrt{n}},$$

avec

$$\mathcal{H}_0 : \tilde{\tau}^d = t \quad \text{vs} \quad \mathcal{H}_1 : \tilde{\tau}^d > t.$$

5.4 Application : test d'indépendance entre des données ordinales multivariées

L'objectif de cette section est de développer un test d'indépendance pour des données ordinales. La procédure sera basée sur le tau de Kendall empirique τ_n^d. Puisque sous l'indépendance multivariée, on a $\tau^d = 0$, un test

d'indépendance dans ce contexte consiste à confronter les hypothèses

$$\mathcal{H}_0 : \tau^d = 0 \quad \text{vs.} \quad \mathcal{H}_1 : \tau^d \neq 0.$$

Du résultat sur la normalité asymptotique de τ_n^d, on rejettera $\mathcal{H}_0$ lorsque

$$\left|\tau_n^d\right| > z_{\alpha/2} \frac{2\sigma_1}{\sqrt{n}}.$$

Une manière d'évaluer l'efficacité d'un test statistique consiste à calculer sa fonction de puissance. Généralement, si Z_n est une statistique dont la valeur critique de rejet est p_α, alors la fonction de puissance associée est définie par

$$\beta = \mathrm{P}\left(Z_n > p_\alpha | \mathcal{H}_1\right).$$

Dans la pratique, il est difficile de calculer la variance σ_1^2. Pour remédier à ce problème, on fera appel à la méthode *jackknife* pour l'estimation d'une variance. Le Jackknife, introduit comme méthode pour l'estimation du biais par Quenouille (1949) [13] et ensuite proposé pour l'estimation de la variance par Tukey (1958), implique la suppression systématique de groupes d'unités, le re-calcul de la statistique avec chaque groupe supprimé tour à tour, et ensuite la combinaison de toutes ces statistiques re-calculées. Le jackknife le plus simple consiste à écarter chaque observation d'un échantillon tour à tour, de re-calculer la statistique pour obtenir des valeurs $\tau_1, \ldots, \tau_n$, et ensuite d'estimer la variance de l'estimateur par

$$\hat{\sigma}_1^2 = \frac{n-1}{n} \sum_{i=1}^{n} \left(\tau_i - \bar{\tau}\right)^2.$$

Dans ce cas, on rejettera $\mathcal{H}_0$ si

$$\left|\tau_n^d\right| > z_{\alpha/2} \frac{\hat{\sigma}_1}{\sqrt{n}}.$$

Dans le tableau suivant sont résumés les résultats de simulations pour tester la performance du test d'indépendance dans le cas bivarié. Dans ce qui suit, on va désigner par α le paramètre de dépendance dans le cas bivarié ainsi que dans le cas multivarié. Les simulations ont été effectuées selon l'algorithme suivant.

- Générer $n = 30$ couples aléatoires suivant des lois de poisson de paramètres m_1, m_2, en effet, on a généré trois variables aléatoires X_1, X_2 et X_3 suivant des lois de poisson de paramètres respectifs $m_1 - \alpha$, $m_2 - \alpha$ et α, on obtient les n couples de variables aléatoires (X, Y) comme suit $X = X_1 + X_3$ et $Y = X_2 + X_3$;
- Effectuer le test d'indépendance basé sur τ_n^d au seuil de signification de 0.05;
- Refaire la même chose $N = 100\ 000$ fois pour estimer la puissance du test selon la formule

$$\hat{\beta} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1} \left(Z_i > Z \left(\alpha \right) \right).$$

Pour les différentes valeurs du couple (m_1, m_2), on remarque que la mesure de la puissance du test de Kendall qu'on a estimé par $\hat{\beta}$ est croissante en fonction du paramètre de dépendance α.

Les valeurs qu'on a eu dans le cas ou $\alpha = 0$ pour les différentes valeurs du couple (m_1, m_2) sont parfaitement cohérentes avec nos hypothèses de départ,

Table 5.2: Puissance de la statistique de Kendall discrète sous des contre-hypothèses de loi de Poisson bivariée

(m_1, m_2)	$\alpha = 0.0$	$\alpha = 0.4$	$\alpha = 0.8$	$\alpha = 1.2$	$\alpha = 1.6$	$\alpha = 2.0$
$(2,2)$	0.058	0.140	0.487	0.883	0.997	1.000
$(4,4)$	0.056	0.075	0.158	0.311	0.519	0.741
$(6,6)$	0.054	0.062	0.100	0.163	0.258	0.382
$(8,8)$	0.055	0.059	0.079	0.110	0.165	0.232
$(10,10)$	0.053	0.058	0.070	0.092	0.122	0.164

ils sont a peu égales ou légèrement différentes de la valeur du risque d'erreur supposé au départ.

Les résultats des simulations obtenus nous poussent a constater que le test d'indépendance qu'on vient de proposer est efficace, il est facile à appliquer contrairement au test de Khi2, où on est amené à construire des classes pour les valeurs, et on a toujours à vérifier des hypothèses ce qui est n'est pas toujours évident.

Dans ce qui suit, on va suivre les mêmes démarches pour tester la validité du test d'indépendance dans le cas multivarié. Par la suite on présente dans le tableau 5.3 le résumé des simulations qui consiste à estimer la puissance du test en fonction du paramètre de dépendance α.

On va illustrer le cas multivarié par le cas de trois variables X, Y et Z, avec $X = X_1 + X_4$, $Y = X_2 + X_4$ et $Z = X_3 + X_4$ ou X_i suivent des lois de poisson de paramètre $m_i - \alpha$, $i = 1, \ldots, 3$ et X_4 suit une loi de poisson de

paramètre α. on va générer $n = 30$ triplets de variables aléatoires et on va suivre le même algorithme pour estimer la fonction de puissance du test du tau de Kendall dans le cas de trois variables tout en faisant varier les valeurs du triplet (m_1, m_2, m_3).

Table 5.3: Puissance de la statistique de Kendall discrète sous des contre-hypothèses de loi de Poisson tri-variée

(m_1, m_2, m_3)	$\alpha = 0.0$	$\alpha = 0.4$	$\alpha = 0.8$	$\alpha = 1.2$	$\alpha = 1.6$	$\alpha = 2.0$
$(2, 2, 2)$	0.065	0.27	0.812	0.992	1.000	1.000
$(4, 4, 4)$	0.062	0.090	0.274	0.559	0.813	0.949
$(6, 6, 6)$	0.062	0.064	0.138	0.275	0.463	0.649
$(8, 8, 8)$	0.061	0.057	0.090	0.163	0.272	0.408
$(10, 10, 10)$	0.062	0.054	0.075	0.121	0.195	0.290

Pour les différentes valeurs du triplet (m_1, m_2, m_3), on remarque que la mesure de la puissance du test de Kendall qu'on a estimé par $\hat{\beta}$ est croissante en fonction du paramètre de dépendance α.

Ces résultats nous poussent à constater que ce test d'indépendance peut être appliqué pour des variables discrètes multivariées, reste à tester la validité pour des combinaisons de variables plus complexes avec d'autres lois de probabilité.

Conclusion

L'étude des mesure de dépendances est de très grande intérêt, il y'a beaucoup de champs d'application, surtout en actuariat, et plus particulièrement pour étudier l'impact de la dépendance des sinistres en assurance IARD (Incendies, Accidents, Risques divers.

Dans ce mémoire, nous avons établis de nouvelles mesures de dépendances dans le cas discret multivarié, des mesures basées sur le tau de Kendall et le Rho de Spearman, ainsi que des propriétés de monotonicité des ses indices. Dans le dernier chapitre, nous avons proposé des tests d'indépendances basés sur le tau de Kendall, nous avons vérifié la validité de ses tests avec le modèle de poisson composé avec choc commun dans le cas bivarié et d-varié avec $d = 3$.

Références

[1] Nelsen, R. B (1999). *An Introduction to Copulas.* Lecture Notes in Statistics no 139. New York : Springer.

[2] Scarsini, M. (1984). "On measures of concordance,". Stochastica 8,

[3] Tchen, A. H., (1980). "Inequalities for distributions with given marginals," The Annals of Probability, 8, 814-827.

[4] Yanagimoto, T., and Okamoto, M (1969). "Partial ordering of permutations and monotonicity of a rank correlation stattistic," Annals of the Institute of Statistical Mathematics, 21, 489-506. 201-218.

[5] Müller, A., and Scarsini, M.(2000). "Some remarks on the supermodular order," Journal of Multivariate Analysis, 73, 107-119.

[6] Capéraà, P., and Genets, C.(1993). "Spearman's ρ is larger than Kendall' τ for positively dependent random varoables," Journal of Nonparametric Statistics, 2, 183-194.

[7] Goodman, L. A., and Kruskal, W. H (1954). "Measures of association for

cross classifications," Journal of the Americain Statistical Association, 49, 732-764.

[8] Shriever, B. F. (1985). Order Dependence, Vrije Universiteit te Amsterdam.

[9] Mesfioui, M., and Tajar, A., and Bouezmarni, T (2006). " A note on concordance measures for discrete data and dependence properties of Poisson model," soumis.

[10] Mesfioui, M., and Tajar, A (2005)."On the properties of some nonparametric concordance measures in the discrete case,". J. Nonparametr. Stat. 17 , no. 5, 541-554.

[11] Cossette, H., and Marceau, E (2000). "The discrete-time risk model with correlated classes of business," Insurance: Mathematics and Economics, Volume 26, Issues 2-3,133-149.

[12] Lee, A. J. (1946). *U-Statistics : theory and practice.* Statistics textbooks and monographs v 109. New York : M. Dekker , c1990.

[13] Quenouille, M. H (1949). On a method of trend elimination. Biometrika 36, 75-91.

[14] Kowalczyk, T., and Niewiadomska-Bugag, M.(1998), "Grade Correspondence Analysis Based on Kendall's tau," in Proceedings of the IFCS-98 Conference, Rome, Italy.