I N S T I T U T D E
S T A T I S T I Q U E

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N
P A P E R

0212

CONSTRAINTS ON CONCORDANCE MEASURES IN BIVARIATE DISCRETE DATA

M. DENUIT and P. LAMBERT

http://www.stat.ucl.ac.be

Constraints on concordance measures in bivariate discrete data

Michel Denuit and Philippe Lambert

Université Catholique de Louvain, Institut de Statistique,
Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium.
Emails: denuit@stat.ucl.ac.be and lambert@stat.ucl.ac.be

March 11, 2002

Abstract

This paper aims to investigate the constraints on dependence measures based on the concept of concordance when discrete random variables are involved. The main technical argument consists in continuing integer-valued random variables by convolution with unit support kernels.

1 Introduction

In the context of dependence, many classical results crucially depend on the continuity assumptions for the marginals. Most papers are devoted to this situation, which is now rather well-known. However, things are less simple when discrete marginals are involved. Many desirable properties of dependence measures no more hold, which can be misleading in practice.

This paper aims to investigate the behavior of dependence measures based on concordance, like Kendall's tau, applied to discrete bivariate data. The main result will be that it becomes senseless to compare the estimated tau to the classical bounds -1 and 1 inherited from the continuous case, since they cannot in general be attained with discrete margins. The population tau is restricted to a narrower margin and this has to be taken into account when its estimated value is interpreted as a measure of the strength of the dependence.

The main theoretical vehicle of this paper is the continuousation of discrete random variables. Specifically, we make those variables continuous by adding a perturbation valued in [0,1]. This approach is very intuitive and allows for easy derivation of interesting results. The main theoretical interest of this construction is that the copula modeling is attractive for continued random variables, whereas it is extremely difficult for the original discrete ones (because of the non-uniqueness of the copula outside the range of the distribution).

We also resort to the concordance order in our analysis. This probabilistic tool compares the strength of dependence expressed by bivariate distributions. It can be regarded as a mathematical translation of the intuitive concept of "being less dependent than" going beyond the simple comparison of Kendall's tau's.

The paper is set out as follows. Section 2 recalls the classical notion of concordance and gives the basic definitions. Section 3 exposes the continuousation principle. It is established there that continuing discrete random variables preserves the concordance order, which is intuitively desirable. Moreover, continuousation leaves Kendall's tau unchanged, a remarkable feature that is exploited to adapt to discrete random variables known results for continuous ones. Section 4 examines the joint distribution of continued random variables. The underlying copula is seen to be the classical bilinear interpolation, playing a central role in the theory. In that respect, the continuousation principle provides an intuitive explanation for this analytical construction. An explicit expression for Kendall's tau is also derived for discrete random variables, in terms of integral transforms expectations. It is also shown on the basis of a simple example that the range (-1,1) of tau in the continuous case is further restricted for discrete outcomes. Section 5 then explores the range of Kendall's tau for discrete marginals. Upper and lower bounds are derived and exemplified for binomial, Poisson and discrete uniform margins. Finally, we end the paper with a discrete analog of the probability integral transformation theorem (which enlightens the difference between continuous and discrete cases) and a discussion.

2 Kendall's tau and discrete responses

2.1 Concordance

The most commonly used measures of association for ordinal variables are those based on the classification of pairs of observations as concordant or discordant. A pair of observations is concordant if the observation with the larger value of X has also the larger value for Y. The pair is discordant if the observation with the larger value of X has the smaller value of Y. If (X_1, Y_1) and (X_2, Y_2) denote independent copies of (X, Y) then the (X_i, Y_i)'s are said to be concordant if $(X_1 - X_2)(Y_1 - Y_2) > 0$ holds true whereas they are said to be discordant when the reverse inequality is valid. Henceforth,

$$\Pr(\text{concordance}) = \Pr[(X_1 - X_2)(Y_1 - Y_2) > 0]$$

and

$$\Pr(\text{discordance}) = \Pr[(X_1 - X_2)(Y_1 - Y_2) < 0].$$

The main difference between the continuous and discrete cases is that in the latter situation many tied pairs (that is, pairs of observations that have equal values of X or equal value of Y) occur in practice. Specifically, if (X_1, Y_1) and (X_2, Y_2) are as above, those pairs are tied if $X_1 = X_2$ or $Y_1 = Y_2$. Henceforth,

$$\Pr(\text{tie}) = \Pr(X_1 = X_2 \text{ or } Y_1 = Y_2).$$

Whereas Kendall's tau is an appreciated tool for measuring the strength of dependence between continuous outcomes, it looses many of its good properties when it is applied to discrete variables. In particular, it is no more distribution-free and its range is restricted to a sub-interval of $[-1, 1]$. Therefore, several dependence measures based on concordance and discordance probabilities have been introduced to deal with discrete random variables. Their respective differences lie in the treatment of ties. Concordance based dependence measures are Goodman's gamma, Kendall's tau b, Stuart's tau c and Somer's Δ. For a general presentation, we refer the interested reader e.g. to Agresti (1996). In this paper, we concentrate our efforts on Kendall's τ. Since all the aforementioned dependence measures are based on concordance and discordance, the reader will adapt the results obtained for Kendall's τ to these other dependence measures.

Before moving to the discussion of Kendall's τ, we briefly recall the well-known copula construction relating a bivariate distribution to its univariate marginals.

2.2 Copula representation for bivariate distributions

Copulas can be used to define bivariate distributions with discrete margins. In opposition to the situation found in the continuous case, there is no unique way to express the joint distribution of two (say) discrete random variables (r.v.'s) as a function of their marginal distributions.

More specifically, denote by X and Y two random variables and by F and G their distribution functions. If H denotes the joint distribution of X and Y, then there exists a distribution C (named *copula*) on $[0,1]^2$ such that

$$H(s,t) = \Pr(X \leq s, Y \leq t) = C(F(s), G(t)) \tag{1}$$

(Sklar, 1959). When the variables are continuous, C is unique. That unicity is only ensured on $\text{range}(F) \times \text{range}(G)$ when the variables are discrete.

2.3 Kendall's tau

Kendall's tau is a widely used measure of dependence between random variables. Let (X,Y) be a couple of random variables with joint distribution $H(s,t)$ that can be related through Equation (1) to the marginal distributions F and G (of X and Y respectively) using a copula C. Assume that (X_1,Y_1) and (X_2,Y_2) are two independent copies of (X,Y). Then, Kendall's tau measures the probability of concordance minus the probability of discordance of (X_1,Y_1) and (X_2,Y_2), i.e.

$$\tau(X,Y) = \Pr(\text{concordance}) - \Pr(\text{discordance}) \tag{2}$$

With continuous random variables, one can show (see e.g. Nelsen, 1998, p. 129) that

$$\begin{aligned}
\tau(X,Y) &= 2\Pr(\text{concordance}) - 1 \\
&= 4\Pr(X_2 \leq X_1, Y_2 \leq Y_1) - 1 \\
&= 4\int\int_{[0,1]^2} C(u,v)dC(u,v) - 1 \qquad (3)
\end{aligned}$$

We see that it is completely determined by the copula and unrelated to the marginal distributions.

When the involved variables X and Y are valued in the non-negative integers,

$$\Pr(\text{concordance}) + \Pr(\text{discordance}) + \Pr(\text{tie}) = 1$$

so that we have

$$\begin{aligned}
\tau(X,Y) &= 2\Pr(\text{concordance}) - 1 + \Pr(\text{tie}) \\
&= 4\Pr(X_2 < X_1, Y_2 < Y_1) - 1 + \Pr(X_1 = X_2 \text{ or } Y_1 = Y_2) \qquad (4)
\end{aligned}$$

since

$$\begin{aligned}
\Pr(X_2 < X_1, Y_2 < Y_1) &= \Pr(X_2 > X_1, Y_2 > Y_1) \\
&= \frac{\Pr(\text{concordance})}{2}.
\end{aligned}$$

As we shall see further in this work, Kendall's tau cannot attain a very large value because a large proportion of the pairs are tied (especially when the number of possible values for X and Y is small).

3 Continuousation of a discrete variable

3.1 Principle

Assume that X is discrete variable valued in a subset $\mathcal{X}$ of the set $\mathbb{N}$ of the non-negative integers, and denote the corresponding non-zero probability masses by

$$f_x = \Pr(X = x),\ x \in \mathcal{X}.$$

We propose to associate to X a continuous random variable X^* such that

$$X^* = X + (U - 1)$$

where U is a continuous random variable valued in (0,1) independent of X with a strictly increasing cdf $L_U(u)$ on $(0, 1)$ sharing no parameter with the distribution of X. We shall say that X *is continued by* U.

Clearly, $X^* \leq X$ a.s. holds. Let $[s]$ be the integer part of $s \in \mathbb{R}$. We also have, for $s \in \mathbb{R}$,

$$\begin{aligned} F^*(s) &= \Pr(X^* \leq s) = \sum_{x \in \mathcal{X} : x \leq [s]} f_x + L_U(s - [s])\ f_{[s+1]} \\ &= F([s]) + L_U(s - [s])\ f_{[s+1]} \end{aligned} \tag{5}$$

and

$$f^*(s) = l_U(s - [s])\ f_{[s+1]} \tag{6}$$

where f^* is the density corresponding to F^* and l_U is the density corresponding to L_U.

Hence, for any $x \in \mathbb{N}$,

$$F(x) = F^*(x)\ ;\ f_x = \int_{x-1}^{x} f^*(s)ds. \tag{7}$$

The constraints on the distribution $L_U(u)$ of U are easy to meet. The most natural choice is certainly a uniform distribution on $(0, 1)$ for U. In that case, for $u \in (0, 1)$, we have

$$L_U(u) = u\ ;\ l_U(u) = 1$$

that can be used in Equations (5) and (6) to specify the distribution and the density of X^*.

In some instances, it might be interesting to consider alternative choices for L_U (as a mixture of uniforms, for instance, in order to recover F as an appropriate limit of F^*).

3.2 Continuousation preserves the concordance order

Intuitively, two random variables X and Y are concordant when large values of X tend to be associated with large values of Y. Several attempts have been made to formulate this concept precisely; see e.g. Tchen (1980) or Scarsini (1984). The problem of comparing the strength of dependence expressed by bivariate distributions is of prime importance for modeling purposes. A formalization of the intuitive idea of "being less dependent than" has been proposed by Yanagimoto and Okamoto (1969); it is the concordance order $\prec_c$ (partially) ranking the bivariate distributions with given univariate margins according to the strength of positive association they express.

This stochastic ordering is defined as follows: given two random vectors (X_1, Y_1) and (X_2, Y_2) with identical marginals, (X_2, Y_2) is said to be more concordant than (X_1, Y_1), denoted as $(X_1, Y_1) \prec_c (X_2, Y_2)$, if

$$\Pr(X_1 \leq s, Y_1 \leq t) \leq \Pr(X_2 \leq s, Y_2 \leq t)$$

holds for all $s, t \in \mathbb{R}$.

If (X_1, Y_1) is the independent version of (X_2, Y_2) then $(X_1, Y_1) \prec_c (X_2, Y_2)$ means that (X_2, Y_2) is positively dependent by quadrants (PQD, in short). The interpretation of this dependence notion stems from the inequality

$$F(s)G(t) \leq \Pr(X_2 \leq s, Y_2 \leq t)$$

valid for all $s, t \in \mathbb{R}^2$. In words, it says that the probability for X_2 and Y_2 to be small (i.e. smaller than some thresholds s and t) is at least as large as it would be were they independent.

Assume now that $(X_1, Y_1) \prec_c (X_2, Y_2)$ holds for some pairs (X_1, Y_1) and (X_2, Y_2) of discrete r.v.'s. If X_1 (Y_1) and X_2 (Y_2) are continued by the same r.v. U (V), with U and V independent, we have

$$\begin{aligned}
&\Pr[X_1^* \leq s, Y_1^* \leq t] = \Pr[X_1 + (U-1) \leq s, Y_1 + (V-1) \leq t] \\
&= \int\int_{u,v\in[0,1]} \Pr[X_1 \leq s-u+1, Y_1 \leq t-v+1]\; l_U(u)\; l_V(v)\; du\; dv \\
&\leq \int\int_{u,v\in[0,1]} \Pr[X_2 \leq s-u+1, Y_2 \leq t-v+1]\; l_U(u)\; l_V(v)\; du\; dv \\
&= \Pr[X_2^* \leq s, Y_2^* \leq t].
\end{aligned}$$

so that we have established the implication

$$(X_1, Y_1) \prec_c (X_2, Y_2) \Rightarrow (X_1^*, Y_1^*) \prec_c (X_2^*, Y_2^*). \tag{8}$$

In particular, if (X, Y) is PQD, then (X^*, Y^*) also is.

Relation (8) ensures the coherence of the continuousation in our context since

(i) continuing random couples does not modify the perception we have about the strength of dependence they express;

(ii) continuing random couples preserves PQD.

3.3 Continuousation preserves Kendall's tau

Since Kendall's tau is based on concordance and discordance probabilities, let us examine how the continuousation affects these probabilities. Let (X_1, Y_1) and (X_2, Y_2) be independent copies of (X, Y). Assume that

- X_i and Y_i are continued by U_i and V_i respectively,
- U_1, U_2, V_1, V_2 are independent,
- U_1 and U_2 (V_1 and V_2) have the same distribution,

Under these conditions, we may write

$$\begin{aligned}
& \overset{*}{\Pr}(\text{concordance}) \\
= \; & \Pr[(X_1^* - X_2^*)(Y_1^* - Y_2^*) > 0] \\
= \; & \Pr[(X_1 + U_1 - X_2 - U_2)(Y_1 + V_1 - Y_2 - V_2) > 0] \\
= \; & \Pr[X_1 = X_2, Y_1 = Y_2] \;\; \Pr[(U_1 - U_2)(V_1 - V_2) > 0] \\
& + \Pr[X_1 = X_2, Y_1 > Y_2] \;\; \Pr[U_1 - U_2 > 0] \\
& + \Pr[X_1 = X_2, Y_1 < Y_2] \;\; \Pr[U_1 - U_2 < 0] \\
& + \Pr[X_1 > X_2, Y_1 = Y_2] \;\; \Pr[V_1 - V_2 > 0] \\
& + \Pr[X_1 < X_2, Y_1 = Y_2] \;\; \Pr[V_1 - V_2 < 0] \\
& + \Pr[(X_1 - X_2)(Y_1 - Y_2) > 0]
\end{aligned}$$

Since $U_1 - U_2$ and $V_1 - V_2$ are continuous r.v.'s with densities symmetric about zero, we get

$$\Pr[(X_1^* - X_2^*)(Y_1^* - Y_2^*) > 0] \;\; = \;\; \frac{1}{2}\Pr(\text{tie}) + \Pr[(X_1 - X_2)(Y_1 - Y_2) > 0],$$

that is,

$$\overset{*}{\Pr}(\text{concordance}) = \Pr(\text{concordance}) + \frac{1}{2}\Pr(\text{tie}).$$

Considering (3)-(4), the latter relation straightly yields

$$\tau(X,Y) = \tau(X^*,Y^*). \tag{9}$$

Yanagimoto and Okamoto (1969) have shown that several dependence measures, including Kendall's tau, are preserved under the concordance order provided the marginal distributions of the random couples were continuous. Tchen (1980, Corollary 3.2) extended this result to discrete marginals with finitely many atoms. In our framework, a similar result is easily deduced from (8), as shown next.

Indeed, assume that X_1 (Y_1) and X_2 (Y_2) are continued using U (V) and that U and V are independent. We have

$$\begin{array}{rcl}(X_1,Y_1) \prec_c (X_2,Y_2) & \overset{(8)}{\Rightarrow} & (X_1^*,Y_1^*) \prec_c (X_2^*,Y_2^*) \\ & \overset{\text{Yanagimoto}}{\Rightarrow} & \tau(X_1^*,Y_1^*) \leq \tau(X_2^*,Y_2^*) \\ & \overset{(9)}{\Leftrightarrow} & \tau(X_1,Y_1) \leq \tau(X_2,Y_2)\end{array}$$

Hence, given two random couples (X_1,Y_1) and (X_2,Y_2) with integer-valued components,

$$(X_1,Y_1) \prec_c (X_2,Y_2) \Rightarrow \tau(X_1,Y_1) \leq \tau(X_2,Y_2). \tag{10}$$

It is worth mentioning that the latter implication holds without any constraints on the size of the supports of X and Y, which slightly improves Tchen's result. Moreover, the inequality between Kendall's τ's is strict when (X_1,Y_1) and (X_2,Y_2) are not identically distributed.

4 Joint distribution of continued variables

4.1 Starred copula

Consider two discrete variables X and Y. They can be continued by (independent) U and V respectively yielding X^* and Y^* with distribution F^* and G^*.

If $H(s,t)$ denotes the joint distribution of X and Y, then there exists a copula C uniquely defined on range$(F) \times$ range(G) such that representation

(1) is valid. The choices made for the distributions of U and V, and the copula C completely determine the joint distribution $H^*(s,t)$ of X^* and Y^*. In order to check this assertion, let z_d denote the fractional part of z (that is, $z_d = z - [z]$). Then, we have

$$\begin{aligned}
H^*(s,t) &= \Pr(X^* \leq s, Y^* \leq t) \\
&= \Pr(X \leq [s], Y \leq [t]) \\
&\quad + L_U(s_d) \Pr(X = [s]+1, Y \leq [t]) \\
&\quad + L_V(t_d) \Pr(X \leq [s], Y = [t]+1) \\
&\quad + L_U(s_d) L_V(t_d) \Pr(X = [s]+1, Y = [t]+1) \\
&= L_U(s_d)\Big\{L_V(t_d)\ C[F([s]+1), G([t]+1)] + [1 - L_V(t_d)]\ C[F([s]+1), G([t])]\Big\} \\
&\quad + (1 - L_U(s_d))\Big\{L_V(t_d)\ C[F([s]), G([t]+1)] + [1 - L_V(t_d)]\ C[F([s]), G([t])]\Big\}
\end{aligned} \tag{11}$$

As $H^*(s,t)$ is the joint distribution of continuous random variables, there exists a unique copula C^* on $[0,1]^2$ such that

$$H^*(s,t) = C^*[F^*(s), G^*(t)] \quad \forall (s,t) \in \mathbb{R}^2$$

given by

$$C^*(u,v) = H^*[F^{*-1}(u), G^{*-1}(v)] \tag{12}$$

First define two inverse functions for the discrete distribution F:

$$\begin{aligned}
\underline{F}^{-1}(u) &= \max\{x \in \mathbb{N} : F(x) \leq u\} \in \mathbb{N} \\
\overline{F}^{-1}(u) &= \min\{x \in \mathbb{N} : F(x) \geq u\} \in \mathbb{N}
\end{aligned}$$

Obviously, $\underline{F}^{-1}(u)$ and $\overline{F}^{-1}(u)$ coincide for u in the range of F. If we set

$$\underline{u}^F = F(\underline{F}^{-1}(u)) \quad ; \quad \overline{u}^F = F(\overline{F}^{-1}(u))$$

then for u outside the range of F,

$$F^{*-1}(u) = \underline{F}^{-1}(u) + L_U^{-1}\left(\frac{u - \underline{u}^F}{\overline{u}^F - \underline{u}^F}\right);$$

the same type of expression can be obtained for $G^{*-1}(v)$. Combining these starred inverses with Equations (11) and (12), we obtain the starred copula

$$\begin{aligned}
C^*(u,v) &= \frac{u - \underline{u}^F}{\overline{u}^F - \underline{u}^F}\left\{\frac{v - \underline{v}^G}{\overline{v}^G - \underline{v}^G} C(\overline{u}^F, \overline{v}^G) + \frac{\overline{v}^G - v}{\overline{v}^G - \underline{v}^G} C(\overline{u}^F, \underline{v}^G)\right\} \\
&+ \frac{\overline{u}^F - u}{\overline{u}^F - \underline{u}^F}\left\{\frac{v - \underline{v}^G}{\overline{v}^G - \underline{v}^G} C(\underline{u}^F, \overline{v}^G) + \frac{\overline{v}^G - v}{\overline{v}^G - \underline{v}^G} C(\underline{u}^F, \underline{v}^G)\right\}
\end{aligned} \tag{13}$$

where (u, v) is outside Range$(F)\times$Range(G). Of course, C and C^* coincide on Range$(F)\times$Range(G). C^* is a bilinear interpolation of the C copula at the surrounding points $\{\underline{u}^F, \overline{u}^F\} \times \{\underline{v}^G, \overline{v}^G\}$ of range(F) $\times$ range(G). We clearly see that the choice made for the distributions of U and V does not influence the starred copula.

It is worth mentioning that the bilinear interpolation (13) has been extensively used in the statistical literature. Our approach can thus be regarded as a probabilistic interpretation of the analytic bilinear interpolation of C.

4.2 Kendall's tau for continued random variables

We can obtain an explicit formula for $\tau(X^*, Y^*)$ (and hence for $\tau(X, Y)$) using (3):

$$\begin{aligned}
\tau(X^*, Y^*) &= 4\int\int_{[0,1]^2} C^*(u,v)dC^*(u,v) - 1 \\
&= 4\sum_{x=0}^{+\infty}\sum_{y=0}^{+\infty}\int_{F_{x-1}}^{F_x}\int_{G_{y-1}}^{G_y} C^*(u,v)dC^*(u,v) - 1
\end{aligned}$$

where

$$F_x = \Pr(X \leq x),\ F_{-1} = 0 \ \ ; \ \ G_y = \Pr(Y \leq y),\ G_{-1} = 0$$

Now, on $(F_{x-1}, F_x) \times (G_{y-1}, G_y)$,

$$dC^*(u,v) = \frac{\partial^2 C^*(u,v)}{\partial u \partial v}dudv = \frac{1}{f_x g_y}\Pr(X = x, Y = y)dudv$$

Hence

$$\begin{aligned}
&\tau(X^*, Y^*) \\
=\ & -1 + 4\sum_{x=0}^{+\infty}\sum_{y=0}^{+\infty}\frac{1}{f_x g_y}\Pr(X = x, Y = y)\int_{F_{x-1}}^{F_x} du \int_{G_{y-1}}^{G_y} dv\ C^*(u,v) \\
=\ & -1 + \sum_{x=0}^{+\infty}\sum_{y=0}^{+\infty}\Pr(X = x, Y = y) \\
& \{C^*(F_x, G_y) + C^*(F_x, G_{y-1}) + C^*(F_{x-1}, G_y) + C^*(F_{x-1}, G_{y-1})\} \\
=\ & \sum_{x=0}^{+\infty}\sum_{y=0}^{+\infty}\Pr(X = x, Y = y) \qquad (14) \\
& \{C(F_x, G_y) + C(F_x, G_{y-1}) + C(F_{x-1}, G_y) + C(F_{x-1}, G_{y-1}) - 1\}
\end{aligned}$$

as C and C^* are equal on range$(F) \times$ range(G). Expressed in terms of r.v.'s, (14) yields

$$\tau(X,Y) = EH(X,Y)+EH(X,Y-1)+EH(X-1,Y)+EH(X-1,Y-1)-1, \tag{15}$$

which can be regarded as a discrete analog of the representation obtained by Schweizer and Wolff (1981) in the continuous case.

4.3 A simple example

Consider a joint model for discrete Bernoulli variables X and Y where failure and success are coded using the integers 0 and 1 respectively. Denote by p_X and p_Y the respective probabilities of success. If X and Y are continued using U and V, and a copula C^* is considered to model the joint distribution of the starred variables, then

$$H^*(s,t) = C^*[F^*(s), G^*(t)]$$

Let h_{ij} denote $\Pr(X = i, Y = j)$, $i, j \in \{0, 1\}$. Using Equation (14), we obtain

$$\tau(X^*, Y^*) = 2[h(0,0)h(1,1) - h(1,0)h(0,1)] \tag{16}$$

which is simply twice the odds ratio (under the considered models for the margins and for the copula). As expected from (9), we obtain the same expression for $\tau(X,Y)$ from a direct application of Equation (4).

Given the marginal probabilities of success p_X and p_Y, we can rewrite Kendall's tau as

$$\tau(X^*, Y^*) \quad = \quad 2[h(1,1) - p_X p_Y] \tag{17}$$

As soon as $h(1,1)$ is fixed, the whole bivariate distribution is specified. That probability of joint success is constrained by

$$\max\{0, p_X + p_Y - 1\} \leq h(1,1) \leq \min\{p_X, p_Y\}$$

which corresponds to the usual Fréchet bounds.

Considering (17), $\tau(X^*, Y^*)$ is minimum when

$$h(1,1) = \max\{0, p_X + p_Y - 1\}$$

in which case

$$\tau(X^*, Y^*) = \begin{cases} -2p_X p_Y & \text{when} \quad p_X + p_Y < 1 \\ -2(1-p_X)(1-p_Y) & \text{when} \quad p_X + p_Y \geq 1 \end{cases} \tag{18}$$

Likewise, $\tau(X^*, Y^*)$ is maximum when

$$h(1,1) = \min\{p_X, p_Y\}$$

in which case

$$\tau(X^*, Y^*) = \begin{cases} 2p_X(1-p_Y) & \text{when} \quad p_X < p_Y \\ 2p_Y(1-p_X) & \text{when} \quad p_X \geq p_Y \end{cases} \tag{19}$$

The lower and upper bounds (18)-(19) directly follow from (10) since the Fréchet lower and upper bounds are the $\prec_c$-minimum and maximum, respectively, once the marginals have been fixed.

The largest possible value for $\tau(X^*, Y^*)$ is obtained when $h(1,1) = h(0,0) = 0.5$ and $h(1,0) = h(0,1) = 0$ (zero probability of discordance) in which case $\tau(X^*, Y^*)$ is equal to 0.50. Similarly, the smallest possible value for $\tau(X^*, Y^*)$ is -0.50 when $h(0,0) = h(1,1) = 0$ (zero probability of concordance) and $h(1,0) = h(0,1) = 0.5$.

Thus, even in the most favorable cases, we see that Kendall's tau cannot reach 1 (-1). We shall come back to this point in Section 5. Let us just give here an intuitive explanation of this fact, together with a rigorous treatment of the case $F = G$. Assume that X and Y are perfectly positively dependent, that is, (X, Y) is distributed as $(F^{-1}(W), G^{-1}(W))$ for some unit uniform r.v. W. The continuousation of these r.v.'s gives

$$(X^*, Y^*) = \Big(F^{-1}(W) + U - 1, G^{-1}(W) + V - 1\Big).$$

Even if X and Y were perfectly dependent, we see that X^* and Y^* are made less dependent because of the addition of independent r.v.'s U and V. Specifically, if the copula C for (X, Y) is $\min\{u, v\}$ on Range$(F)\times$Range(G), the copula C^* for (X^*, Y^*) given in (13) does not coincide with the Fréchet upper bound everywhere in the unit square.

It is easy to check these assertions in the particular case $F = G$. In this case,

$$(X^*, Y^*) \prec_c (F^{*-1}(W), F^{*-1}(W)) \Rightarrow \tau(X^*, Y^*) < \tau(F^{*-1}(W), F^{*-1}(W)) = 1.$$

In this case, (X^*, Y^*) and $(F^{*-1}(W), F^{*-1}(W))$ cannot be identically distributed since $X^* - Y^* = U - V \neq 0$ a.s.

5 Kendall's tau upper bound for discrete variables

5.1 A first lower and upper bound for Kendall's tau

Before deriving a stricter upper bound for Kendall's tau, let us first examine Equation (2) to obtain a maximal upper bound for Kendall's tau. Obviously, $\tau(X, Y)$ will be maximal when the probability of discordance is zero (which is not always possible). Hence, since

$$\Pr(\text{concordance}) + \Pr(\text{discordance}) + \Pr(\text{tie}) = 1$$

we conclude, after combination with Equation (4), that

$$\tau(X, Y) \leq 1 - \Pr(\text{tie})$$

with equality when the probability of discordance is zero.

Similarly, $\tau(X, Y)$ will be minimal when the probability of concordance is zero (which is not always possible) yielding

$$\tau(X, Y) \geq -1 + \Pr(\text{tie})$$

with equality when the probability of concordance is zero.

Let (X_1, Y_1) and (X_2, Y_2) be independent copies of (X, Y). Combining these last results with

$$\Pr(\text{tie}) \quad = \quad \Pr(X_1 = X_2) + \Pr(Y_1 = Y_2) - \Pr(X_1 = X_2, Y_1 = Y_2)$$

we conclude that

$$\begin{aligned} -1 \quad < \quad & -1 + \max\{\Pr(X_1 = X_2), \Pr(Y_1 = Y_2)\} \\ & \leq -1 + \Pr(\text{tie}) \leq \tau(X, Y) \leq 1 - \Pr(\text{tie}) \\ & \leq 1 - \max\{\Pr(X_1 = X_2), \Pr(Y_1 = Y_2)\} < 1 \end{aligned} \tag{20}$$

whatever the joint distribution of X and Y. Let us mention that the inequalities in (20) can also be derived from the representation (15).

The inequalities in (20) seem rather surprising because we know from (9) that continuing X and Y does not modify Kendall's tau. Nevertheless, we have to keep in mind that even if the joint distribution of (X, Y) is the Fréchet upper bound, X^* and Y^* are not perfectly dependent so that their Kendall's tau is strictly less than one.

Binomial margins: assume that

$$X \sim \text{Bin}(n, p_X) \;\; ; \;\; Y \sim \text{Bin}(n, p_Y)$$

with $n > 1$.

We have computed the (non optimal) upper bound for $\tau(X, Y)$ based on Equation (20) when

$$X \sim \text{Bin}(5, p_X) \;\; ; \;\; Y \sim \text{Bin}(5, p_Y)$$

for a grid of (p_X, p_Y) values in $\mathcal{P}_X \times \mathcal{P}_Y$ where

$$\mathcal{P}_X = \{0, 0.01, \ldots, 1.00\} \;\; ; \;\; \mathcal{P}_Y = \{0, 0.05, \ldots, 0.50\}$$

These upper bounds are displayed in Figure 1 where each curve joins the values of these bounds for a given value of p_Y and $p_X \in \mathcal{P}_X$, higher curves corresponding to larger values of p_Y.

5.2 Stricter bounds for Kendall's tau

5.2.1 General case: arbitrary discrete margins

The knowledge of the copula C joining the discrete variables X and Y can be used to derive stricter bounds for Kendall's tau. A first way to use that information is to compute exactly the probability of a tie given by

$$\begin{aligned} \Pr(\text{tie}) &= \Pr(X_1 = X_2) + \Pr(Y_1 = Y_2) - \Pr(X_1 = X_2, Y_1 = Y_2) \\ &= \sum_{x \in \mathcal{X}} f_x^2 + \sum_{y \in \mathcal{Y}} g_y^2 \\ &\quad - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} [C(F_x, G_y) + C(F_{x-1}, G_{y-1}) - C(F_x, G_{y-1}) - C(F_{x-1}, G_y)] \end{aligned}$$

and to inject it in Equation (20).

As second way consists of using Equation (14). We shall focus on the upper bound. However, similar work can be done to derive a lower bound for Kendall's tau by substituting the Fréchet lower bound copula to the Fréchet upper bound copula in the text below.

We know that

$$\tau(X, Y) \leq \tau(F^{-1}(W), G^{-1}(W)).$$

Further,

$$\begin{aligned} \tau(F^{-1}(W), G^{-1}(W)) &= \Pr[(F^{-1}(W) - F^{-1}(W'))(G^{-1}(W) - G^{-1}(W')) > 0] \\ &\quad - \Pr[(F^{-1}(W) - F^{-1}(W'))(G^{-1}(W) - G^{-1}(W')) < 0] \end{aligned}$$

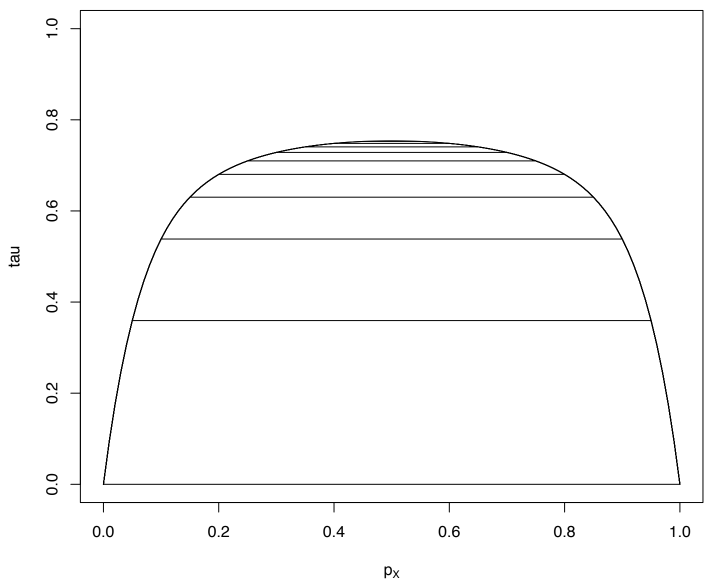

Figure 1: Kendall's tau (non optimal) upper bound when $X \sim \text{Bin}(5, p_X)$ and $Y \sim \text{Bin}(5, p_Y)$ with $(p_X, p_Y) \in \{0, 0.01, \ldots, 1.00\} \times \{0, 0.05, \ldots, 0.50\}$, higher curves corresponding to higher values of p_Y

where W and W' are independent unit uniform r.v.'s. Clearly, the second term vanishes and

$$\tau(F^{-1}(W), G^{-1}(W)) = 2\,\text{Pr}[F^{-1}(W) < F^{-1}(W'), G^{-1}(W) < G^{-1}(W')].$$

Since

$$\begin{aligned} \tau(X, X) &= 2\,\text{Pr}[F^{-1}(W) < F^{-1}(W')] \\ \tau(Y, Y) &= 2\,\text{Pr}[G^{-1}(W) < G^{-1}(W')] \end{aligned}$$

we clearly have that

$$\tau(X, Y) \leq \min\Big\{\tau(X, X), \tau(Y, Y)\Big\}.$$

Invoking Equation (15) yields

$$\tau(X, Y) \leq \min\left\{\sum_{x=0}^{+\infty} f_x(4F_{x-1} + f_x - 1), \sum_{y=0}^{+\infty} g_y(4G_{y-1} + g_y - 1)\right\} \qquad (21)$$

providing a possibly stricter upper bound for Kendall's tau. That bound is optimal when X and Y share the same distribution.

Poisson margins: the upper bound (21) for Kendall's tau corresponding to identical Poisson margins related by the Fréchet upper bound copula is displayed for values of the mean parameter up to 50 on Figure 2. As the probability of ties tends to zero when μ tends to infinity, the upper bound for Kendall's tau tends to 1 as expected.

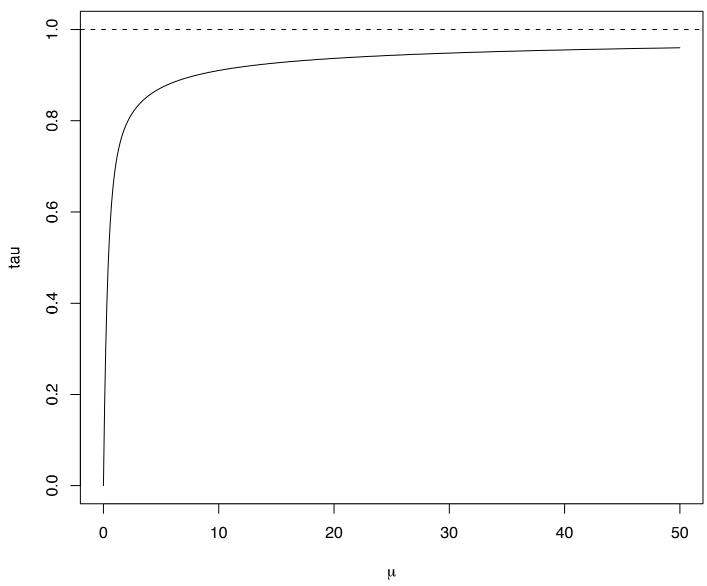

Figure 2: Kendall's tau upper bound (21) when both random variables are Poisson with mean μ.

Binomial margins: we have computed $\tau_{\max}(X, Y)$ when

$$X \sim \text{Bin}(5, p_X) \quad ; \quad Y \sim \text{Bin}(5, p_Y)$$

for a grid of (p_X, p_Y) values in $\mathcal{P}_X \times \mathcal{P}_Y$ where

$$\mathcal{P}_X = \{0, 0.01, \ldots, 1.00\} \quad ; \quad \mathcal{P}_Y = \{0, 0.05, \ldots, 0.50\}$$

The values of $\tau_{\max}(X, Y)$ are displayed in Figure 3 where each curve joins the values of $\tau_{\max}$ for a given value of p_Y and $p_X \in \mathcal{P}_X$, higher curves corresponding to larger values of p_Y.

These upper bounds can be compared to the (non optimal) bounds for $\tau(X,Y)$ displayed in Figure 1.

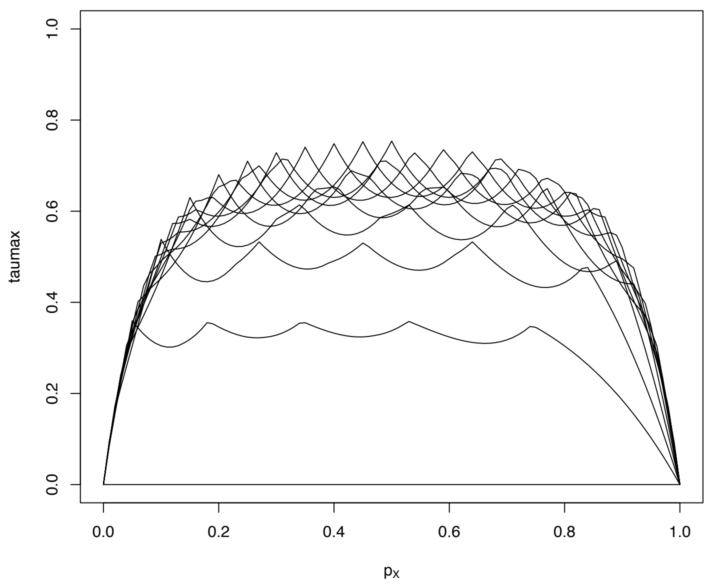

Figure 3: Kendall's tau upper bound (21) when $X \sim \text{Bin}(5, p_X)$ and $Y \sim \text{Bin}(5, p_Y)$ with $(p_X, p_Y) \in \{0, 0.01, \ldots, 1.00\} \times \{0, 0.05, \ldots, 0.50\}$, higher curves corresponding to higher values of p_Y

5.2.2 Special case: identically distributed discrete margins with finite number of atoms

Consider a discrete distribution F with a finite domain

$$\mathcal{X} = \{0, 1, \ldots, N-1\}$$

Equation (21) becomes

$$\tau(X,Y) \;\; \leq \;\; \sum_{x=0}^{N-1} f_x(4F_x - 3f_{x-1} - 1)$$

We can bound that expression (independently of the true underlying discrete distribution) by noting that it is maximum when the probability mass is equally distributed within the N cells, i.e. when

$$f_x = \frac{1}{N}$$

Indeed, let us define the function

$$S_N = \sum_{x=0}^{N-1} f_x \left(4 \sum_{y=0}^{x} f_y - 3f_x - 1 \right).$$

The latter can be split into

$$\begin{aligned} S_N &= \sum_{x=0}^{N-2} f_x \left(4 \sum_{y=0}^{x} f_y - 3f_x - 1 \right) + 3 \sum_{x=0}^{N-2} f_x - 3 \left(\sum_{x=0}^{N-2} f_x \right)^2 \\ &= 4 \sum_{x=0}^{N-2} f_x \sum_{y=0}^{x} f_y - 3 \sum_{x=0}^{N-2} f_x^2 + 2 \sum_{x=0}^{N-2} f_x - 3 \left(\sum_{x=0}^{N-2} f_x \right)^2. \end{aligned}$$

Now, the partial derivative of $S_N$with respect to f_0 gives

$$\frac{\partial}{\partial f_0} S_N = -4f_0 - 2 \sum_{x=1}^{N-2} f_x + 2$$

so that

$$\frac{\partial}{\partial f_0} S_N = 0 \Leftrightarrow f_0 = \frac{1 - \sum_{x=1}^{N-2} f_x}{2} = \frac{f_0 + f_{N-1}}{2} \Leftrightarrow f_0 = f_{N-1}.$$

A similar reasoning yields $f_i = f_{N-1}$ for all i, whence the announced result follows.

We thus have

$$\tau(X,Y) \leq \sum_{x=0}^{N-1} \frac{1}{N} \left(4\frac{x}{N} + \frac{1}{N} - 1 \right) = 1 - \frac{1}{N}$$

with equality achieved if and only if

$$f_x = \frac{1}{N}$$

Bernoulli margins: we recover the 0.5 upper bound that is realized when the probabilities of success are both equal to 0.5

Binomial margins: assume that

$$X \sim \mathrm{Bin}(n,p) \quad ; \quad Y \sim \mathrm{Bin}(n,p)$$

with $n > 1$. Then, the above result claims that

$$\tau(X,Y) < 1 - \frac{1}{n+1}$$

with a strict inequality because a binomial distribution with $n > 1$ can never be uniform.

6 Discrete analog of the PITT

The classical Probability Integral Transformation Theorem (PITT, in short) ensures that for any random variable Z with continuous distribution function G, the random variable $G(Z)$ is uniformly distributed on $[0,1]$. This result plays a central role in many statistical procedures, but is no more true when G has some discontinuities, like with discrete random variables.

Let us first relate $F(X)$ to the unit uniform distribution for counting random variables X. Since

$$\begin{aligned} \Pr(F(X) \leq p) &= \Pr(X \leq \underline{F}^{-1}(p)) \\ &= F(\underline{F}^{-1}(p)) \leq p \end{aligned}$$

for any $p \in (0,1)$, we have

$$\Pr(F(X) \leq p) \leq p \qquad \forall p \in (0,1)$$

In particular, $EF(X) > 1/2$. This indicates that the graph of the distribution function of the r.v. $F(X)$ is always below that of the unit uniform distribution. Therefore, $F(X)$ tends to be larger than a unit uniform r.v. (because the probability for $F(X)$ to be small – i.e. below some threshold p – is lower than the corresponding probability computed under the unit uniform distribution).

Let us now establish a discrete analog of the PITT for counting random variables. If X is a discrete random variable X with domain $\mathcal{X}$ in $\mathbb{N}$ (see Section 3) such that

$$f_x = \Pr(X = x),\ x \in \mathcal{X},$$

continued by U, then

$$F^*(X^*) = F^*[X + (U-1)] = F([X^*]) + f_{[X^*]+1}U = F(X-1) + f_X U$$

is uniformly distributed on $(0,1)$.

7 Discussion

One motivation for this work was generated by Vandenhende and Lambert (2000)[1] where the dependence between longitudinal ordinal data was modeled using copulas. It was pointed that Kendall's tau-b could not solely be calculated from the copula (dependence) parameter, but also required consideration of the marginal distributions.

More generally, when the margins are discrete, the strength of dependence cannot be assessed solely by inspecting Kendall's tau (or modified versions of Kendall's tau like tau-b): such measures should be evaluated knowing the extremal values that they can take. Here, we provide formulas for these values using the continuouosation argument.

That argument provides an elegant tool that can be used further to translate, in a multivariate discrete setting, copula based results that are only valid when working with multivariate continuous responses.

Acknowledgements

Most of this paper was written while P. Lambert was visiting the Applied Statistics Department of the University of Reading with the support of a grant of the Fonds National de la Recherche Scientifique (FNRS, Belgium).

References

Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* New York: Wiley & Sons.

Nelsen, R. (1998). *An Introduction to Copulas.* New York: Springer-Verlag.

Scarsini, M. (1984). On measures of concordance. *Stochastica 8*, 201–218.

Schweizer, B. and E. Wolff (1981). On nonparametric measures of dependence for random variables. *The Annals of Statistics 9*, 879–885.

Sklar, A. (1959). Fonctions de répartitions à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris 8*, 229–231.

Tchen, A. (1980). Inequalities for distributions with given marginals. *The Annals of Probability 8*, 814–827.

[1]ftp://www.stat.ucl.ac.be/pub/papers/dp/dp00/dp0025.ps

Vandenhende, F. and P. Lambert (2000). Modeling repeated ordered categorical data using copulas. *Discussion Paper 00-25, Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.*

Yanagimoto, T. and M. Okamoto (1969). Partial orderings of permutations and monotonicity of a rank correlation statistic. *Ann. Inst. Statist. Math. 21*, 489–506.