Contents

Submitted to the Bernoulli

A consistent test of independence based on a sign covariance related to Kendall's tau

WICHER BERGSMA and ANGELOS DASSIOS

London School of Economics and Political Science,
Houghton Street,
London WC2A 2AE,
United Kingdom,
E-mail: w.p.bergsma@lse.ac.uk*;* a.dassios@lse.ac.uk

The most popular ways to test for independence of two ordinal random variables are by means of Kendall's tau and Spearman's rho. However, such tests are not consistent, only having power for alternatives with 'monotonic' association. In this paper we introduce a natural extension of Kendall's tau, called τ^*, which is nonnegative and zero if and only if independence holds, thus leading to a consistent independence test. Furthermore, normalization gives a rank correlation which can be used as a measure of dependence, taking values between zero and one. A comparison with alternative measures of dependence for ordinal random variables is given, and it is shown that, in a well-defined sense, τ^* is the simplest, similarly to Kendall's tau being the simplest of ordinal measures of monotone association. Simulation studies show our test compares well with the alternatives in terms of average p-values.

AMS 2000 subject classifications: Primary 62H20, 62H15; secondary 62G10.
Keywords: measure of association, concordance, discordance, sign test, ordinal data, permutation test, copula.

1. Introduction

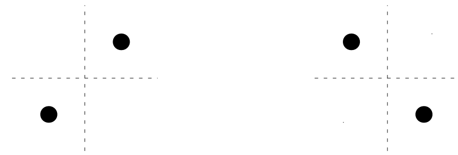
A random variable X is called *ordinal* if its possible values have an ordering, but no distance is assigned to pairs of outcomes. Ordinal variables may be continuous, categorical, or mixed continuous/categorical. Ordinal data frequently arise in many fields, though especially often in social and biomedical science (Kendall & Gibbons, 1990; Agresti, 2010). Ordinal data methods are also often applied to real-valued (interval level) data in order to achieve robustness.

The two most popular measures of association for ordinal random variables X and Y are Kendall's tau (τ) (Kendall, 1938) and Spearman's rho (ρ_S) (Spearman, 1904), which may be defined as

$$\tau = E\,\mathrm{sign}[(X_1 - X_2)(Y_1 - Y_2)] \qquad \rho_S = 3E\,\mathrm{sign}[(X_1 - X_2)(Y_1 - Y_3)]$$

where the (X_i, Y_i) are independent replications of (X, Y) (Kruskal, 1958). The factor 3 in the expression for ρ_S occurs to obtain a measure whose range is $[-1, 1]$. From the definitions, probabilistic interpretations of τ and ρ_S can be derived. Firstly,

$$\tau = \Pi_{C_2} - \Pi_{D_2} \tag{1}$$



(a) Concordant pair (b) Discordant pair

Figure 1. Concordant and discordant pairs of points associated with Kendall's tau

(a) Concordant triples (b) Discordant triples

Figure 2. Concordant and discordant triples of points associated with Spearman's rho

where Π_{C_2} is the probability that two observations are concordant and Π_{D_2} the probability that they are discordant (see Figure 1). Secondly,

$$\rho_S = \Pi_{C_3} - \Pi_{D_3}$$

where Π_{C_3} is the probability that three observations are concordant and Π_{D_3} the probability that they are discordant (see Figure 2). It can be seen that τ is simpler than ρ_S, in the sense that it can be defined using only two rather than three independent replications of (X, Y), or, more specifically, in terms of probabilities of concordance and discordance of two rather than three points. This was a reason for Kruskal to prefer τ to ρ_S (Kruskal, 1958, page 846).

An alternative definition of ρ_S, which was originally given by Spearman, is as a Pearson correlation between uniform rank scores of the X and Y variables. For continuous random variables, both this and the aforementioned definition lead to the same quantity. However, with this definition, ρ_S is to some extent an *ad hoc* measure, since the choice of scores is arbitrary, and alternative scores (e.g., normal scores) might be used.

A test of independence based on iid data can be obtained by application of the permutation test to an estimator of τ or ρ_S, which is easy to implement and fast to carry out with modern computers. Such ordinal tests are also used as a robust alternative to tests based on the Pearson correlation.

A drawback for certain applications is that τ and ρ_S may be zero even if there is an association between X and Y, so tests based on them are inconsistent for the alternative of a general association. For this reason alternative coefficients have been devised. The best known of these are those introduced by Hoeffding (1948) and Blum, Kiefer, and Rosenblatt (1961). With F_{12} the joint distribution function of (X, Y), and F_1 and F_2 the marginal distribution functions of X resp. Y, Hoeffding's coefficient is given as

$$H = \int [F_{12}(x, y) - F_1(x)F_2(y)]^2 dF_{12}(x, y) \tag{2}$$

and the Blum-Kiefer-Rosenblatt (henceforth: BKR) coefficient as

$$D = \int [F_{12}(x, y) - F_1(x)F_2(y)]^2 dF_1(x)dF_2(y) \tag{3}$$

Both can be seen to be nonnegative with equality to zero under independence. Furthermore, $D = 0$ can also be shown to imply independence. However, the Hoeffding coefficient has a severe drawback, namely that it may be zero even if there is an association, i.e., it does not lead to a consistent independence test. An example is the case that $P(X = 0, Y = 1) = P(X = 1, Y = 0) = 1/2$ (Hoeffding, 1948, page 548).

A third option, especially suitable for categorical data, is the Pearson chi-square test; it is directly applicable to categorical data and can be used for continuous data after a suitable categorization. However, the chi-square test does not take the ordinal nature of the data into account, leading to potential power loss for 'ordinal' alternatives; effectively the chi-square test treats the data as nominal rather than ordinal (see also Agresti, 2010).

Although H and D have simple mathematical formulas, they seem to be rather arbitrary, and many variants are possible (see also Section 3.4). For this reason we decided to develop a probabilistic interpretation of H and D (given in Section 3 of this paper). However, we then noticed that H and D were unnecessarily complex, and that a clearly simpler and natural alternative coefficient was possible. Our new coefficient is a direct modification of Kendall's τ, which we call τ^*. It is nonnegative and zero if and only if independence holds. Like τ and ρ_S, we show that H, D and τ^* equal the difference of concordance and discordance probabilities of a number of independent replications of (X, Y). Analogously to the aforementioned way that τ is simpler than ρ_S, τ^* is simpler than D and H in that only four independent replications of (X, Y) are required, whereas H needs five, and D needs six. It appears to us that relative simplicity of interpretation of a coefficient is of utmost importance, and that this is also the main reason for the current popularity of Kendall's tau. In particular, when it was introduced in the pre-computer age in 1938, the sample value of Kendall's tau was much harder to compute than the sample value of Spearman's rho, which had been in use since 1904 (Kruskal, 1958). In spite of this, judging by the number of Google Scholar hits, both currently appear to be about equally popular[1].

The organization of this paper is as follows. In Section 2, we first define τ^*, and then state our main theorem that $\tau^* \geq 0$ with equality if and only if independence holds. Furthermore, we provide a probabilistic interpretation in terms of concordance and discordance probabilities of four points. Section 5 contains the proof of the main theorem. The proof turns out to be surprisingly involved for such a simple to formulate coefficient, and the ideas in the proof may be useful for other related research. A comparison with the Hoeffding and the BKR coefficients is given in Section 3, and new probabilistic interpretations for these coefficients are given. In Section 4 we give a description of independence testing via the permutation test and a simulation study compares average p-values of our test and the aforementioned other two tests. Our test compares well with the other two in this respect.

2. Definition of τ^* and statement of its properties

We denote iid sample values by $(x_1, y_1), \ldots, (x_n, y_n)$, but will also use $\{(X_i, Y_i)\}$ to denote iid replications of (X, Y) in order to define population coefficients. The empirical value t of Kendall's

[1]The Google Scholar search `"kendall's tau" OR "kendall tau"` gave us 15,200 hits and the search `"spearman's rho" OR "spearman rho"` 17,500

tau is

$$t = \frac{1}{n^2} \sum_{i,j=1}^{n} \operatorname{sign}(x_i - x_j) \operatorname{sign}(y_i - y_j)$$

and its population version is

$$\tau = E \operatorname{sign}(X_1 - X_2) \operatorname{sign}(Y_1 - Y_2)$$

(Kruskal, 1958; Kendall & Gibbons, 1990). With

$$\begin{aligned} s(z_1, z_2, z_3, z_4) &= \operatorname{sign}(z_1 - z_4)(z_3 - z_2) \\ &= \operatorname{sign}(|z_1 - z_2|^2 + |z_3 - z_4|^2 - |z_1 - z_3|^2 - |z_2 - z_4|^2) \end{aligned}$$

we obtain

$$t^2 = \frac{1}{n^4} \sum_{i,j,k,l=1}^{n} s(x_i, x_j, x_k, x_l) s(y_i, y_j, y_k, y_l)$$

and

$$\tau^2 = E s(X_1, X_2, X_3, X_4) s(Y_1, Y_2, Y_3, Y_4)$$

Replacing squared differences in s by absolute values of differences, we define

$$a(z_1, z_2, z_3, z_4) = \operatorname{sign}\left(|z_1 - z_2| + |z_3 - z_4| - |z_1 - z_3| - |z_2 - z_4|\right) \tag{4}$$

This leads to a modified version of t^2,

$$t^* = \frac{1}{n^4} \sum_{i,j,k,l=1}^{n} a(x_i, x_j, x_k, x_l) a(y_i, y_j, y_k, y_l)$$

and the corresponding population coefficient

$$\tau^* = \tau^*(X, Y) = E a(X_1, X_2, X_3, X_4) a(Y_1, Y_2, Y_3, Y_4)$$

The quantities t^* and τ^* are new, and the main result of the paper is the following:

Theorem 1 *It holds true that $\tau^*(X, Y) \geq 0$ with equality if and only if X and Y are independent.*

The proof is given in Section 5.

If the sign functions are omitted from τ^*, we obtain the covariance introduced by Bergsma (2006) and Székely, Rizzo, and Bakirov (2007). They showed that for arbitrary real random variables X and Y, this covariance is nonnegative with equality to zero if and only if X and Y are independent.

By the Cauchy-Schwarz inequality, the normalized value

$$\tau_b^* = \frac{\tau^*(X, Y)}{\sqrt{\tau^*(X, X) \tau^*(Y, Y)}}$$

does not exceed one. (Note that this notation is in line with Kendall's τ_b, defined analogously.)

The definition of τ^* can easily be extended to X and Y in arbitrary metric spaces, but unfortunately Theorem 1 does not extend then, as it is possible that $\tau^* < 0$. This is shown by the following example. Consider a set of points $\{u_1, \ldots, u_8\} \subset \mathbf{R}^8$, where $u_i = (u_{i1}, \ldots, u_{i8})'$ such that $u_{ii} = 3$, $u_{ij} = -1$ if $i \neq j$ and $i, j \leq 4$ or $i, j \geq 5$, and $u_{ij} = 0$ otherwise. Suppose Y is uniformly distributed on $\{0, 1\}$, and given $Y = 0$, X is uniformly distributed on $u_1, \ldots, u_4$, and given $Y = 1$, X is uniformly distributed on $u_5, \ldots, u_8$. Then $\tau^* = -1/64$.

Note that $\tau^*(X, Y)$ is a function of the copula, which is the joint distribution of $F_1(X)$ and $F_2(Y)$, where F_1 and F_2 are the cumulative distribution functions of X and Y. Nelsen (2006, Chapter 5) explores the way in which copulas can be used in the study of dependence between random variables, paying particular attention to Kendall's tau and Spearman's rho.

We now give a probabilistic interpretation of τ^*. Recall that Kendall's tau is the probability that a pair of points is concordant minus the probability that a pair of points is discordant. Our τ^* is proportional to the probability that two pairs are 'jointly' concordant, plus the probability that two pairs are 'jointly' discordant, minus the probability that, 'jointly', one pair is discordant and the other concordant. Here, 'jointly' refers to there being a common axis separating the two points of each of the two pairs.

To use a slightly different terminology which will be convenient, we say that a set of four points is concordant if two pairs are either 'jointly' concordant or 'jointly' discordant, while four points are called discordant if, 'jointly', one pair is concordant and the other is discordant. These configurations are given in Figure 3. In mathematical notation, a set of four points $\{(x_1, y_1), \ldots, (x_4, y_4)\}$ is concordant if there is a permutation (i, j, k, l) of $(1, 2, 3, 4)$ such that

$$(x_i, x_j < x_k, x_l) \& \left[(y_i, y_j < y_k, y_l) || (y_i, y_j > y_k, y_l)\right]$$

and discordant if there is a permutation (i, j, k, l) of $(1, 2, 3, 4)$ such that

$$\left[(x_i, x_j < x_k, x_l) || (x_i, x_j > x_k, x_l)\right] \& \left[(y_i, y_k < y_j, y_l) || (y_i, y_k > y_j, y_l)\right]$$

where $||$ and $\&$ are logical OR resp. AND, and $I(z_1, z_2 < z_3, z_4)$ is shorthand for $I(z_1 < z_3 \& z_1 < z_4 \& z_2 < z_3 \& z_2 < z_4)$. It is straightforward to verify that

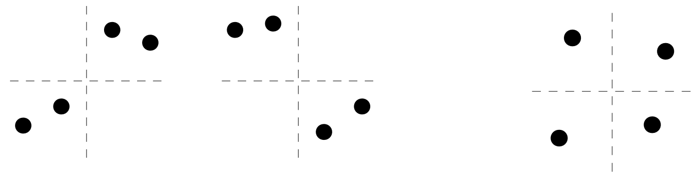
$$\begin{aligned} a(z_1, z_2, z_3, z_4) &= I(z_1, z_3 < z_2, z_4) + I(z_1, z_3 > z_2, z_4) \\ &\quad - I(z_1, z_2 < z_3, z_4) - I(z_3, z_4 > z_1, z_2) \end{aligned}$$

where I is the indicator function. Hence,

$$\begin{aligned} \tau^* &= 4P(X_1, X_2 < X_3, X_4 \& Y_1, Y_2 < Y_3, Y_4) + \\ &\quad 4P(X_1, X_2 < X_3, X_4 \& Y_1, Y_2 > Y_3, Y_4) - \\ &\quad 8P(X_1, X_2 < X_3, X_4 \& Y_1, Y_3 < Y_2, Y_4) \end{aligned} \tag{5}$$

Denoting the probability that four randomly chosen points are concordant as Π_{C_4} and the probability that they are discordant as Π_{D_4}, we obtain that the sum of the first two probabilities on the right hand side of (5) equals $\Pi_{C_4}/6$, while the last probability equals $\Pi_{D_4}/24$. Hence,

$$\tau^* = \frac{2\Pi_{C_4} - \Pi_{D_4}}{3} \tag{6}$$



(a) Concordant points (b) Discordant points

Figure 3. Configurations of concordant and discordant quadruples of points associated with τ^*. The dotted axes indicate strict separation of points in different quadrants; within a quadrant, no restrictions apply on the relative positions of points.

It can be seen that t^* and τ^* do not depend on the scale at which the variables are measured, but only on the ranks or grades of the observations. Four points are said to be *tied* if they are neither concordant nor discordant. Clearly, for continuous distributions the probability of tied observations is zero. Hence, under independence, when all configurations are equally likely, $\Pi_{C_4} = 1/3$ and $\Pi_{D_4} = 2/3$, and if one variable is a strictly monotone function of the other, then $\Pi_{C_4} = 1$ and $\Pi_{D_4} = 0$.

3. Comparison to other tests

The two most popular (almost) consistent tests of independence for ordinal random variables are those based on Hoeffding's H and BKR's D, given in (2) and (3). Probabilistic interpretations for these coefficients do not appear to have been given in the literature, and the present section gives these. The probabilistic interpretation shows that τ^* is simpler than both H and D. Since $H = 0$ does not imply independence if the distributions are discrete, it should perhaps not be used, and we are left with two coefficients, τ^* and D, of which τ^* is the simplest. Further discussions of ordinal data and nonparametric methods for independence testing are given Agresti (2010), Hollander and Wolfe (1999) and Sheskin (2007).

3.1. Hoeffding's *H*

Hoeffding's (1948) coefficient for measuring deviation from independence for a bivariate distribution function is given by (2) (see also Blum et al., 1961; Hollander & Wolfe, 1999 and Wilding & Mudholkar, 2008). An alternative formulation given by Hoeffding is

$$H = \frac{1}{4} E\phi(X_1, X_2, X_3)\phi(X_1, X_4, X_5)\phi(Y_1, Y_2, Y_3)\phi(Y_1, Y_4, Y_5)$$

where $\phi(z_1, z_2, z_3) = I(z_1 \geq z_2) - I(z_1 \geq z_3)$. A severe drawback of Hoeffding's H is that, for discrete distributions, it is not necessarily zero under independence. An example is the case that $P(X = 0, Y = 1) = P(X = 1, Y = 0) = 1/2$ (Hoeffding, 1948, page 548).

Interestingly, Hoeffding's H has an interpretation in terms of concordance and discordance probabilities closely related to the interpretation of τ^*. With

$$\begin{aligned}
F_{12}(x,y) &= P(X \le x, Y \le y) \\
F_{1\bar{2}}(x,y) &= P(X \le x, Y > y) = F_1(x) - F_{12}(x,y) \\
F_{\bar{1}2}(x,y) &= P(X > x, Y \le y) = F_2(y) - F_{12}(x,y) \\
F_{\overline{12}}(x,y) &= P(X > x, Y > y) = 1 - F_1(x) - F_2(y) + F_{12}(x,y)
\end{aligned}$$

we have the equality

$$F_{12} - F_1 F_2 = F_{12} F_{\overline{12}} - F_{1\bar{2}} F_{\bar{1}2} \tag{7}$$

Let five points be H-concordant if four are configured as in Figure 3(a) and the fifth is on the point where the axes cross and, analogously, five points are H-discordant if four are configured as in Figure 3(b) and the fifth is on the point where the axes cross. Denote the probabilities of H-concordance and discordance by Π_{C_5} and Π_{D_5}. Then, omitting the arguments x and y,

$$\int \left(F_{12}^2 F_{\overline{12}}^2 + F_{1\bar{2}}^2 F_{\bar{1}2}^2\right) dF_{12} = \frac{2!2!1!}{5!}\Pi_{C_5} = \frac{1}{30}\Pi_{C_5}$$

and

$$\int F_{12} F_{1\bar{2}} F_{\bar{1}2} F_{\overline{12}} dF_{12} = \frac{1}{5!}\Pi_{D_5} = \frac{1}{120}\Pi_{D_5}$$

Hence, using (7),

$$H = \int \left(F_{12} F_{\overline{12}} - F_{1\bar{2}} F_{\bar{1}2}\right)^2 dF_{12} = \frac{2\Pi_{C_5} - \Pi_{D_5}}{60}$$

3.2. The Blum-Kiefer-Rosenblatt coefficient D

The coefficient D is given by (3), and tests based on it were first studied by Blum et al. (1961). Analogously to Hoeffding's H, a probabilistic interpretation of D can be given but based on six rather than five independent replications of (X,Y). In particular, let six points be D-concordant if four are configured as in Figure 3(a), such that a fifth point lies on the x-axis and a sixth on the y-axis. Analogously, six points are D-discordant if four are configured as in Figure 3(b), such that a fifth point lies on the x-axis and a sixth on the y-axis. Denote the probabilities of D-concordance and discordance by Π_{C_6} and Π_{D_6}. Then, omitting the arguments x and y,

$$\int \left(F_{12}^2 F_{\overline{12}}^2 + F_{1\bar{2}}^2 F_{\bar{1}2}^2\right) dF_1 dF_2 = \frac{2!2!2!}{6!}\Pi_{C_6} = \frac{1}{90}\Pi_{C_6}$$

and

$$\int F_{12} F_{1\bar{2}} F_{\bar{1}2} F_{\overline{12}} dF_1 dF_2 = \frac{2!}{6!}\Pi_{D_6} = \frac{1}{360}\Pi_{D_6}$$

Hence, using (7),

$$D = \int \left(F_{12} F_{\overline{12}} - F_{1\bar{2}} F_{\bar{1}2}\right)^2 dF_1 dF_2 = \frac{2\Pi_{C_6} - \Pi_{D_6}}{180}$$

It follows from results in Bergsma (2006) that in the continuous case, with

$$h(z_1, z_2, z_3, z_4) = |z_1 - z_2| + |z_3 - z_4| - |z_1 - z_3| - |z_2 - z_4|,$$

$$D = Eh(F_1(X_1), F_1(X_2), F_1(X_3), F_1(X_4))h(F_2(Y_1), F_2(Y_2), F_2(Y_3), F_2(Y_4))$$

A similar formulation was given by Feuerverger (1993), who used characteristic functions for its derivation. This connection of Feuerverger's work to that of Blum et al. does not appear to have been noted before.

Replacing absolute values in h by squares, it is straightforward to show that D reduces to Spearman's rho.

3.3. Comparison of τ^*, H, and D

We can see that Hoeffding's H has two drawbacks compared to τ^*. Firstly, it is more complex in that it is based on concordance and discordance of five points rather than four and, secondly, it is not necessarily zero under independence for discrete distributions. D has one drawback compared to τ^*, in that it is more complex in that it is based on concordance and discordance of six points rather than four. Following Kruskal's (1958) preference for Kendall's tau over Spearman's rho due to its relative simplicity, the same preference might be expressed for τ^* compared to D.

3.4. Related work

If one of the variables is binary, our approach leads to the Cramér von Mises test, as shown in Section 3 in Dassios and Bergsma (2012).

We now describe further approaches to obtaining consistent independence tests for ordinal variables described in the literature. It may be noted that H and D are special cases of a general family of coefficients, which can be formulated as

$$Q_{g,h} = Q_{g,h}(X, Y) = \int g(|F_{12}(x, y) - F_1(x)F_2(y)|)d[h(F_{12})(x, y)] \tag{8}$$

For appropriately chosen g and h, $Q_{g,h} = 0$ if and only if X and Y are independent. Instances were studied by De Wet (1980), Deheuvels (1981), Schweizer and Wolff (1981) and Feuerverger (1993) (where the former two focussed on asymptotic distributions of empirical versions, while the latter two focussed on population coefficients). Alternatively, Rényi (1959) proposed *maximal correlation*, defined as

$$\rho^+ = \sup_{g,h} \rho(g(X), h(Y))$$

where the supremum is taken over square integrable functions. Though applicable to ordinal random variables, ρ^+ does not utilize the ordinal nature of the variables. Furthermore, it is hard to estimate, and has the drawback that it may equal one for distributions arbitrarily 'close' to independence (Kimeldorf & Sampson, 1978). An ordinal variant, proposed by Kimeldorf and Sampson (1978), was to maximize the correlation over nondecreasing square integrable functions.

		Y						
		1	2	3	4	5	6	7
X	1	2	1	0	0	0	1	2
	2	1	2	0	0	0	2	1
	3	0	0	2	1	2	0	0
	4	0	0	1	1	1	0	0
	5	0	0	1	2	1	0	0

Table 1. Artificial contingency table containing multinomial counts. Permutation tests based on Kendall's tau and the Pearson chi-square statistic do not yield a significant association ($p = .99$ resp. $p = .25$), but a permutation test based on t^* yields $p = 0.035$

		Change in size of Ulcer Crater (Y)			
		Larger	Healed ($< \frac{2}{3}$)	Healed ($\geq \frac{2}{3}$)	Healed
Treatment group (X)	A	6	4	10	12
	B	11	8	8	5

Table 2. Results of study comparing two treatments of gastric ulcer

4. Testing independence

A suitable test for independence is a permutation test which rejects the independence hypothesis for large values of t^*. For every permutation π of the observed y-values, the sample τ^*-value t^*_π is computed, and the p-value is the proportion of the $\{t^*_\pi\}$ which exceed t^*. As is well-known, the permutation test conditions on the empirical marginal distributions, which are sufficient statistics for the independence model. In practice, the number of permutations may be too large to compute and a random sample of permutations is taken, which is also called a resampling test. Note that there doesn't seem to be a need for an asymptotic approximation to the sampling distribution of t^*.

Direct evaluation of t^* requires computational time $O(n^4)$, which may be practically infeasible for moderately large samples, but t^* can be well-approximated by taking a random sample of subsets of four observations. The proof of Theorem 1 suggests that the complexity can be reduced to $O(n^3)$. An open problem is what the minimum computational complexity of computing t^* is.

Below, we compare various tests of independence using an artificial and a real data set and via a simulation study.

4.1. Examples

An artificial multinomial table of counts is given in Table 1, where X and Y are ordinal variables with 5 and 7 categories. Visually, we can detect an association pattern, but as it is non-monotonic a test based on Kendall's tau does not yield a significant p-value. The chi-square test also yields a non-significant $p = 0.252$, while a permutation test based on t^* yields $p = 0.032$, giving evidence of an association. We also did tests based on D, which yields $p = 0.047$, and the test based on Hoeffding's H yields $p = 0.028$. In this example, using a consistent test designed for ordinal data, evidence for an association can be found, which is not possible with a nominal data test like the chi-square test or with a test based on Kendall's tau. For all tests except Hoeffding's 10^6 resamples were used, and for Hoeffding's test 4,000 resamples were used.

Table 2 shows data from a randomized study to compare two treatments for a gastric ulcer crater, and was previously analyzed in Agresti (2010). Using 10^5 resamples, the chi-square test

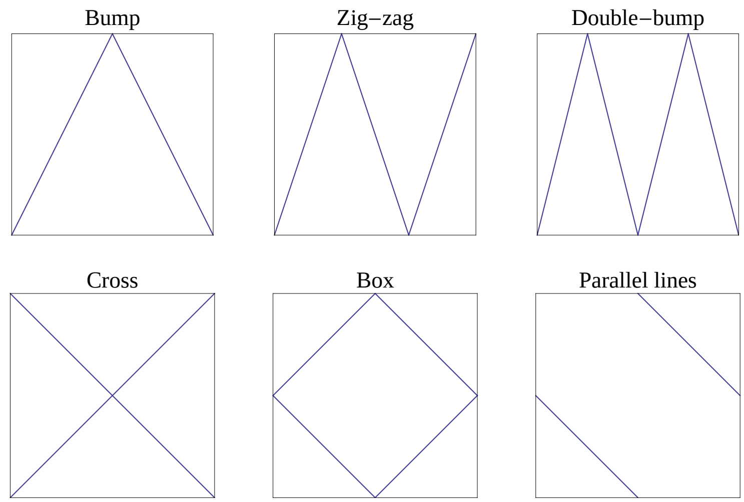


Figure 4. Simulations were done for data generated from the uniform distribution on the lines within each of the six boxes. For all except the Zig-zag and the Parallel lines, the ordinary correlation is zero.

yields $p = 0.118$, Kendall's tau yields $p = 0.019$, t^* yields $p = 0.028$, D yields $p = 0.026$, and using 10^4 resamples Hoeffding's H yields $p = 0.006$.

4.2. Simulated average p-values for independence tests based on D, H, and τ^*

Any of the three tests can be expected to have most power of the three for certain alternatives, and least power of the three for others. Given the broadness of possible alternatives, it cannot be hoped to get a simple description of alternatives for which any single test is the most powerful. However, some insight may be gained by looking at average p-values for a set of carefully selected alternatives.

In Figure 4, six boxes with lines in them are represented, and we simulated from the uniform distribution on these lines. The first five maximize or minimize the correlation between some simple orthogonal functions for given uniform marginals. In particular, say the boxes represent the square $[0,1] \times [0,1]$, then the Bump, Zig-zag and Double bump distributions maximize, for given uniform marginals,

$$\rho[\cos(2\pi X), \cos(\pi Y)], \;\; \rho[\cos(3\pi X), \cos(\pi Y)], \;\text{ and } \rho[\cos(4\pi X), \cos(\pi Y)]$$

respectively. The Cross and Box distributions respectively maximize and minimize, for given uniform marginals,

$$\rho[\cos(2\pi X), \cos(2\pi Y)]$$

As they represent in this sense extreme forms of association, these distributions should yield good insight in the comparative performance of the tests. Furthermore, the Parallel lines distribution

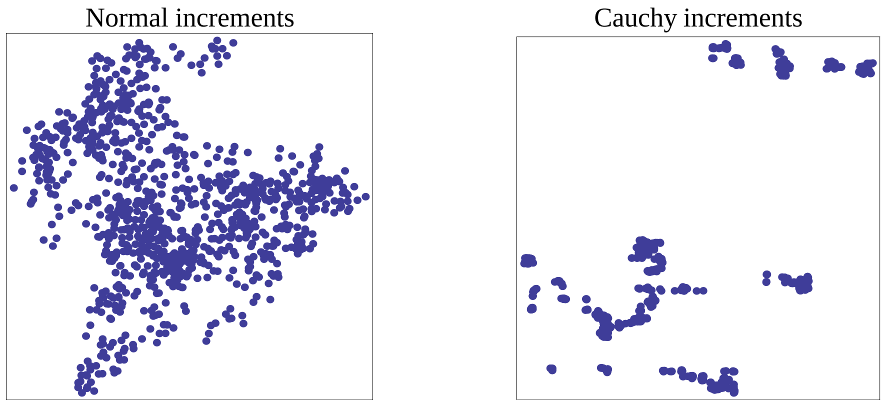


Figure 5. 1000 points of a random walk. In the first plot the (x, y) increments are independent normals, in the second they are independent Cauchy variables.

was chosen because it is simple and demonstrates a weakness of Hoeffding's test, as it has comparatively very little power here (we did not manage to find a distribution where D or τ^* fare so comparatively poorly). Note that all six distributions have uniform marginals and so are copulas, and several were also discussed in Nelsen (2006)

We also did a Bayesian simulation, based on random distributions with dependence. In particular, the data are $(X_1, Y_1), \ldots, (X_n, Y_n)$, where, for iid $(\varepsilon_{1i}, \varepsilon_{2i})$,

$$\begin{aligned}(X_1, Y_1) &= (\varepsilon_{11}, \varepsilon_{21}) \\ (X_{i+1}, Y_{i+1}) &= (X_i, Y_i) + (\varepsilon_{1i}, \varepsilon_{2i}) \qquad i = 1, \ldots, n-1\end{aligned}$$

Of course, the (X_i, Y_i) are not iid, but conditioning on the empirical marginals the permutations of the Y-values give equally likely data sets under the null hypothesis of independence, so the permutation test is valid. Two distributions for the increments $(\varepsilon_{1i}, \varepsilon_{2i})$ were used: independent normals and independent Cauchy distributions. In Figure 5, points generated in this way are plotted. Note that for the Cauchy increments, the heavy tails of the marginal distributions are automatically taken care of by the use of ranks, so in that respect the three tests described here are particularly suitable.

Finally, we also simulated normally distributed data with correlation 0.5.

Average p-values are given in Table 3, where all averages are over at least 40,000 simulations (for D, we did 200,000 simulations). Hoeffding's test compares extremely badly with our test for the parallel lines distribution, and is worse than our test for the random walks, but outperforms our test for the Zig-zag, Double-bump, Cross and Box distributions. The reason for the poor performance of Hoeffding's test for the parallel lines distribution is that five points can only be concordant (see Section 3.1) if they all lie on a single line (a discordant set of five points has zero probability). Similarly, for the Zig-zag, Double-bump and Cross concordant sets of five points can be seen to be especially likely, so these choices of distributions favour the Hoeffding test.

Distribution	Sample size n	Average p-value		
		D	H	τ^*
Random walk (normal increments)	50	.061	.080	.061
Random walk (Cauchy increments)	30	.039	.065	.031
Bump	12	.087	.061	.045
Zig-zag	25	.083	.011	.036
Double-bump	30	.056	.005	.019
Cross	50	.052	.003	.021
Box	50	.070	.008	.019
Parallel lines	10	.055	.710	.076
Normal distribution ($\rho = .5$)	30	.055	.052	.073

Table 3. Average p-values. See Figures 4 and 5 and the text for explanations.

Note that Hoeffding's test is less suitable for general use because it is not necessarily zero under independence if there is a positive probability of tied observations.

The BKR test fares slightly worse than ours for the random walk with Cauchy increments, and significantly worse than ours for the Bump, Zig-zag, Cross and Box distributions, and does somewhat better than ours for the normal distribution. It appears that the BKR test has more power than ours for a monotone alternative (such as the normal distribution), at the cost of less power for some more complex alternatives.

5. Proof of Theorem 1

Here we give the proof of Theorem 1 for arbitrary real random variables X and Y. A shorter proof for continuous X and Y is given by Dassios and Bergsma (2012). Readers wishing to gain an understanding of the essence of the proof may wish to study the shorter proof first.

First consider three real valued random variables U, $V$and W. They have continuous densities $\tilde{f}(x)$, $\tilde{g}(x)$ and $\tilde{k}(x)$ as well as probability masses $f(x_i)$, $g(x_i)$ and $k(x_i)$ at points $x_1, x_2, \ldots$. We also define

$$F(x) = P(U < x) = \sum_{x_i < x} f(x_i) + \int_{y<x} \tilde{f}(y)\,\mathrm{dy},$$

$$G(x) = P(V < x) = \sum_{x_i < x} g(x_i) + \int_{y<x} \tilde{g}(y)\,\mathrm{dy}$$

and

$$K(x) = P(W < x) = \sum_{x_i < x} k(x_i) + \int_{y<x} \tilde{k}(y)\,\mathrm{dy}\,.$$

We will also use $H(x) = \frac{K(x)}{G(x)}$. Note that $H(x)$ also admits the representation

$$H(x) = \sum_{x_i < x} h(x_i) + \int_{y<x} \tilde{h}(y)\,\mathrm{dy}\,.$$

but unlike the other three function that are non-decreasing $\tilde{h}(x)$ and $h(x_i)$ can take negative values.

We start by proving the following intermediate result.

Lemma 1 *Assume that $H(x) > 0$ implies $G(x) > 0$. Define*

$$A = 2\sum (F(x_i) - G(x_i))(F(x_i)g(x_i) - G(x_i)f(x_i))\frac{K(x_i)}{G^2(x_i)} -$$

$$\sum (F(x_i)g(x_i) - G(x_i)f(x_i))^2 \frac{K(x_i)}{G^2(x_i)} +$$

$$2\int (F(x) - G(x))\left(F(x)\tilde{g}(x) - G(x)\tilde{f}(x)\right)\frac{K(x)}{G^2(x)}\,\mathrm{dx}$$

where summation is over all x_i such that $H(x_i) > 0$ and integration over all x such that $H(x) > 0$.

We then have $A \geq 0$ with equality iff $F \equiv G$ (the two distributions are identical).

Proof: We can rewrite

$$A = 2\sum (F(x_i) - G(x_i))(F(x_i)g(x_i) - G(x_i)f(x_i))\frac{H(x_i)}{G(x_i)} -$$

$$\sum (F(x_i)g(x_i) - G(x_i)f(x_i))^2 \frac{H(x_i)}{G(x_i)} +$$

$$2\int (F(x) - G(x))\left(F(x)\tilde{g}(x) - G(x)\tilde{f}(x)\right)\frac{H(x)}{G(x)}\,\mathrm{dx}$$

For simplicity we denote $F(x), G(x), H(x), f(x_i), g(x_i), H(x_i), \tilde{f}(x), \tilde{g}(x)$ and $\tilde{h}(x)$ by $F, G, H, f, g, h, \tilde{f}, \tilde{g}$ and $\tilde{h}$. We have

$$A = 2\sum (F - G)((F - G)g - G(f - g))\frac{H}{G} +$$

$$2\int (F - G)\left((F - G)\tilde{g} - G\left(\tilde{f} - \tilde{g}\right)\right)\frac{H}{G}\,\mathrm{dx} -$$

$$\sum ((F - G)g - G(f - g))^2 \frac{H}{G} =$$

$$2\sum (F - G)^2 \frac{H}{G} g + 2\int (F - G)^2 \frac{H}{G}\tilde{g}\,\mathrm{dx} -$$

$$2\sum H(F - G)(f - g) - 2\int H(F - G)\left(\tilde{f} - \tilde{g}\right)\mathrm{dx} -$$

$$\sum ((F - G)g - G(f - g))^2 \frac{H}{G} \tag{9}$$

The function $H(F - G)^2$ vanishes at $-\infty$ and $-\infty$. Considering its integral and sum representation we have

$$2\sum H(F - G)(f - g) + 2\int H(F - G)(f - g)\,\mathrm{dx} +$$

$$\sum (F - G)^2 h + \int (F - G)^2 \tilde{h}\,\mathrm{dx} +$$

$$+2\sum\left(F-G\right)\left(f-g\right)h+\sum\left(f-g\right)^{2}h+\sum H\left(f-g\right)^{2}=0$$

and therefore

$$-2\sum H\left(F-G\right)\left(f-g\right)-2\int H\left(F-G\right)\left(f-g\right)\mathrm{dx}=$$

$$\sum\left(F-G\right)^{2}h+\int\left(F-G\right)^{2}\tilde{h}\,\mathrm{dx}+$$

$$+2\sum\left(F-G\right)\left(f-g\right)h+\sum\left(f-g\right)^{2}h+\sum H\left(f-g\right)^{2}=0 \tag{10}$$

Moreover,

$$\frac{H}{G}\left(\left(F-G\right)g-G\left(f-g\right)\right)^{2}=$$

$$\left(F-G\right)^{2}g^{2}\frac{H}{G}+GH\left(f-g\right)^{2}-2\left(F-G\right)\left(f-g\right)Hg. \tag{11}$$

Substituting (10) and (11) into (9), and denoting $M=F-G$, $m=f-g$ and $\tilde{m}=\tilde{f}-\tilde{g}$ we have

$$A=\sum M^{2}\left(2g\frac{H}{G}+h-g^{2}\frac{H}{G}\right)+2\sum Mm\left(h+gH\right)+\sum m^{2}\left(H+h-GH\right)+$$

$$\int M^{2}\left(2\tilde{g}\frac{H}{G}+\tilde{h}\right)\mathrm{dx}=$$

$$\sum\left(M+m\right)^{2}\left(g\frac{H}{G+g}+h\right)+\sum M^{2}\left(2g\frac{H}{G}-g\frac{H}{G+g}-g^{2}\frac{H}{G}\right)-$$

$$2\sum Mm\left(g\frac{H}{G+g}-gH\right)+m^{2}\left(H-GH-g\frac{H}{G+g}\right)+$$

$$\int M^{2}\left(\tilde{g}\frac{H}{G}+\tilde{h}\right)\mathrm{dx}+\int M^{2}\tilde{g}\frac{H}{G}\,\mathrm{dx}=$$

$$\sum\left(M+m\right)^{2}\left(g\frac{H}{G+g}+h\right)+\int M^{2}\left(\tilde{g}\frac{H}{G}+\tilde{h}\right)\mathrm{dx}+\int M^{2}\tilde{g}\frac{H}{G}\,\mathrm{dx}+$$

$$\sum M^{2}\left(g\frac{H}{G}+g^{2}\frac{H\left(1-G-g\right)}{G\left(G+g\right)}\right)-2\sum Mm\left(g\frac{H\left(1-G-g\right)}{G+g}\right)+$$

$$\sum m^{2}\frac{H}{G+g}\left(\left(1-G\right)G+g\left(1-G-g\right)\right).$$

Observe now that since $K=HG$

$$g\frac{H}{G+g}+h=\frac{gH+hG+hg}{G+g}=\frac{k}{G+g}\geq 0$$

and

$$\tilde{g}\frac{H}{G}+\tilde{h}=\frac{\tilde{k}}{G}\geq 0.$$

Moreover the quadratic form

$$M^2\left(g\frac{H}{G}+g^2\frac{H\left(1-G-g\right)}{G\left(G+g\right)}\right)-2Mm\left(g\frac{H\left(1-G-g\right)}{G+g}\right)+$$

$$m^2\frac{H}{G+g}\left(\left(1-G\right)G+g\left(1-G-g\right)\right)$$

is non-negative as we can see that its discriminant is non-positive; this is because

$$g^2\frac{H^2\left(1-G-g\right)^2}{\left(G+g\right)^2}-\left(g\frac{H}{G}+g^2\frac{H\left(1-G-g\right)}{G\left(G+g\right)}\right)\frac{H}{G+g}\left(\left(1-G\right)G+g\left(1-G-g\right)\right)\le$$

$$g^2\frac{H^2\left(1-G-g\right)^2}{\left(G+g\right)^2}-g^2\frac{H^2\left(1-G-g\right)}{G^2}=$$

$$g^2H^2\left(1-G-g\right)\left(\frac{1-G-g}{\left(G+g\right)^2}-\frac{1}{G^2}\right)\le g^2H^2\left(1-G-g\right)\left(\frac{1}{\left(G+g\right)^2}-\frac{1}{G^2}\right)\le 0.$$

All terms in A are non-negative and are equal to zero iff $M\equiv 0$, that is the two distributions F and G are identical. □

Before we prove Theorem 1, we will prove another result as it will be used repeatedly.

Lemma 2 *Let A, B and Cbe events in the same probability space as the random variable X and define*

$$L\left(x^{(1)},x^{(2)}\right)=\left(P\left(A\middle|X=x^{(1)}\right)-P\left(A\middle|X<x^{(1)}\wedge x^{(2)}\right)\right)\cdot$$

$$\left(P\left(A\middle|X=x^{(2)}\right)-P\left(A\middle|X<x^{(1)}\wedge x^{(2)}\right)\right)\cdot$$

$$P\left(B\middle|X<x^{(1)}\wedge x^{(2)}\right)P\left(C\middle|X<x^{(1)}\wedge x^{(2)}\right)\left(P\left(X<x^{(1)}\wedge x^{(2)}\right)\right)^2.$$

We then have

$$E\left(L\left(X_1,X_2\right)\right)\ge 0$$

with equality iff X is independent of the event A.

Proof: Let X have continuous density $\tilde{g}\left(x\right)$ and probability masses $g\left(x_i\right)$ at points $x_1,x_2,\ldots$ and let X have continuous density $\tilde{g}_A\left(x\right)$ and probability masses $g_A\left(x_i\right)$ at points $x_1,x_2,\ldots$ conditionally on $Y\in A$. Define also

$$G\left(x\right)=P\left(X<x\right)=\sum_{x_i<x}g\left(x_i\right)+\int_{y<x}\tilde{g}\left(y\right)\mathrm{dy}$$

and

$$G_A\left(x\right)=P\left(X<x\mid A\right)=\sum_{x_i<x}g_A\left(x_i\right)+\int_{y<x}\tilde{g}_A\left(y\right)\mathrm{dy}\,.$$

Conditioning on values of $X_1 \wedge X_2$ and using Bayes' theorem, we can see that

$$E\left(L\left(X_1,X_2\right)\right)=\left(P\left(A\right)\right)^2\sum P\left(B\left|X<x_i\right.\right)P\left(C\left|X<x_i\right.\right)\cdot$$
$$\left\{2\left(\left(1-G_A\left(x_i\right)\right)G\left(x_i\right)-\left(1-G\left(x_i\right)\right)G_A\left(x_i\right)\right)\left(g_A\left(x_i\right)G\left(x_i\right)-g\left(x_i\right)G_A\left(x_i\right)\right)-\right.$$
$$\left.\left(g_A\left(x_i\right)G\left(x_i\right)-g\left(x_i\right)G_A\left(x_i\right)\right)^2\right\}+$$
$$\left(P\left(Y\in A\right)\right)^2\int P\left(B\left|X<x\right.\right)P\left(C\left|X<x\right.\right)\cdot$$
$$\left(\left(1-G_A\left(x\right)\right)G\left(x\right)-\left(1-G\left(x\right)\right)G_A\left(x\right)\right)\left(\tilde g_A\left(x\right)G\left(x\right)-\tilde g\left(x\right)G_A\left(x\right)\right)\mathrm{dx}=$$
$$P\left(B\right)P\left(C\right)\left(P\left(A\right)\right)^2\sum\frac{K\left(x_i\right)}{G^2\left(x_i\right)}\cdot$$
$$\left\{2\left(G\left(x_i\right)-G_A\left(x_i\right)\right)\left(g_A\left(x_i\right)G\left(x_i\right)-g\left(x_i\right)G_A\left(x_i\right)\right)-\left(g_A\left(x_i\right)G\left(x_i\right)-g\left(x_i\right)G_A\left(x_i\right)\right)^2\right\}+$$
$$P\left(B\right)P\left(C\right)\left(P\left(A\right)\right)^2\int\frac{K\left(x\right)}{G^2\left(x\right)}2\left(G\left(x\right)-G_A\left(x\right)\right)\left(\tilde g_A\left(x\right)G\left(x\right)-\tilde g\left(x\right)G_A\left(x\right)\right)\mathrm{dx},$$

where

$$K\left(x\right)=P\left(X<x\left|B\right.\right)P\left(X<x\left|C\right.\right).$$

The result then follows from Lemma 1 ($F = G_A$). □

Proof of Theorem 1: We need to prove that

$$P\left(Y_1\wedge Y_2>Y_3\vee Y_4,X_3\vee X_4<X_1\wedge X_2\right)+$$
$$P\left(Y_1\wedge Y_2<Y_3\vee Y_4,X_3\vee X_4<X_1\wedge X_2\right)-$$
$$P\left(Y_1\wedge Y_3>Y_2\vee Y_4,X_3\vee X_4<X_1\wedge X_2\right)-$$
$$P\left(Y_1\wedge Y_3<Y_2\vee Y_4,X_3\vee X_4<X_1\wedge X_2\right)\geq 0$$

with equality in the independence case.

Let (X,Y) represent any of the pairs $(X_{i,}Y_i)$. Define now $F_1\left(y\right)=P\left(Y<y\left|X=x^{(1)}\right.\right)$, $F_2\left(y\right)=P\left(Y<y\left|X=x^{(2)}\right.\right)$ and $G\left(y\right)=P\left(Y<y\left|X<x^{(1)}\wedge x^{(2)}\right.\right)$ with the representations

$$F_1\left(x\right)=\sum_{y_i<y}f_1\left(y_i\right)+\int_{z<y}\widetilde{f_1}\left(z\right)\mathrm{dz},$$

$$F_2\left(x\right)=\sum_{y_i<y}f_2\left(y_i\right)+\int_{z<y}\widetilde{f_2}\left(z\right)\mathrm{dz}$$

and

$$G\left(x\right)=\sum_{y_i<y}g\left(y_i\right)+\int_{z<y}\tilde g\left(z\right)\mathrm{dz}\,.$$

Note that conditionally on the event $\Theta = \left\{X_1 = x^{(1)}, X_2 = x^{(2)}, X_3 < x^{(1)} \wedge x^{(2)}, X_4 < x^{(1)} \wedge x^{(2)}\right\}$ the distribution of the minimum of Y_1 and Y_2 has density $(1 - F_1)\,\widetilde{f_2} + (1 - F_2)\,\widetilde{f_1}$ and probability masses $(1 - F_1)\,f_2 + (1 - F_2)\,f_1 - f_1 f_2$ at $x_1, x_2, \ldots$, the distribution of the minimum of Y_3 and Y_4 has density $2\,(1 - G)\,\tilde{g}$ and probability masses $2\,(1 - G)\,g - g^2$, the distribution of the minimum of Y_1 and Y_3 has density $(1 - F_1)\,\tilde{g} + (1 - G)\,\widetilde{f_1}$ and probability masses $(1 - F_1)\,g + (1 - G)\,f_1 - f_1 g$ and the distribution of the minimum of Y_2 and Y_4 has density $(1 - F_2)\,\tilde{g} + (1 - G)\,\widetilde{f_2}$ and probability masses $(1 - F_2)\,g + (1 - G)\,f_2 - f_2 g$. We therefore have (suppressing the arguments of the functions)

$$P\left(Y_1 \wedge Y_2 > Y_3 \vee Y_4\,|\Theta\,\right) + P\left(Y_1 \wedge Y_2 < Y_3 \vee Y_4\,|\Theta\,\right) -$$

$$P\left(Y_1 \wedge Y_3 > Y_2 \vee Y_4\,|\Theta\right) - P\left(Y_1 \wedge Y_3 > Y_2 \vee Y_4\,|\Theta\,\right) =$$

$$\sum \left((1 - F_1)\,f_2 + (1 - F_2)\,f_1 - f_1 f_2\right) G^2 + \sum \left(2\,(1 - G)\,g - g^2\right) F_1 F_2 -$$

$$\sum \left((1 - F_1)\,g + (1 - G)\,f_1 - f_1 g\right) F_2 G - \sum \left((1 - F_2)\,g + (1 - G)\,f_2 - f_2 g\right) F_1 G +$$

$$\int \left((1 - F_1)\,\widetilde{f_2} + (1 - F_2)\,\widetilde{f_1}\right) G^2\,\mathrm{dy} + \int \left(2\,(1 - G)\,g - g^2\right) F_1 F_2\,\mathrm{dy} -$$

$$\int \left((1 - F_1)\,\tilde{g} + (1 - G)\,\widetilde{f_1}\right) F_2 G\,\mathrm{dy} - \int \left((1 - F_2)\,\tilde{g} + (1 - G)\,\widetilde{f_2}\right) F_1 G\,\mathrm{dy} =$$

$$\sum (F_1 - G)\,(F_2 g - G f_2) + \sum (F_2 - G)\,(F_1 g - G f_1) - \sum (F_1 g - G f_1)\,(F_2 g - G f_2) +$$

$$\int (F_1 - G)\left(F_2 \tilde{g} - G \widetilde{f_2}\right) \mathrm{dy} + \int (F_2 - G)\left(F_1 \tilde{g} - G \tilde{f_1}\right) \mathrm{dy} =$$

$$2 \sum (F_1 - G)\,(F_2 - G)\,g - \sum (F_1 - G)\,(f_2 - g)\,G - \sum (F_2 - G)\,(f_1 - g)\,G -$$

$$\sum (F_1 g - G f_1)\,(F_2 g - G f_2) + 2 \int (F_1 - G)\,(F_2 - G)\,\tilde{g}\,\mathrm{dy} -$$

$$\int (F_1 - G)\left(\tilde{f_2} - \tilde{g}\right) \mathrm{Gdy} - \int (F_2 - G)\left(\tilde{f_1} - \tilde{g}\right) \mathrm{Gdy}$$

The function $G\,(F_1 - G)\,(F_2 - G)$ vanishes at $-\infty$ and $-\infty$. Considering its integral and sum representation we have

$$-\sum (F_1 - G)\,(f_2 - g)\,G - \sum (F_2 - G)\,(f_1 - g)\,G -$$

$$\int (F_1 - G)\left(\tilde{f_2} - \tilde{g}\right) \mathrm{Gdy} - \int (F_2 - G)\left(\tilde{f_1} - \tilde{g}\right) \mathrm{Gdy} =$$

$$\sum (F_1 - G)\,(F_2 - G)\,g + \sum (F_1 - G)\,(f_2 - g)\,g + \sum (F_2 - G)\,(f_1 - g)\,g +$$

$$\sum (f_1 - g)\,(f_2 - g)\,G - \sum (f_2 - g)\,(f_1 - g)\,g + \int (F_1 - G)\,(F_2 - G)\,\tilde{g}\,\mathrm{dy} =$$

$$\sum (F_1 + f_1 - G - g)\,(F_2 + f_2 - G - g)\,g +$$

$$\sum\left(f_1-g\right)\left(f_2-g\right)G+\int\left(F_1-G\right)\left(F_2-G\right)\tilde{g}\,\mathrm{dy}\,. \tag{12}$$

Moreover,

$$\begin{gathered}
\left(F_1g-Gf_1\right)\left(F_2g-Gf_2\right)=\left(F_1-G\right)\left(F_2-G\right)g^2-\left(F_1-G\right)\left(f_2-g\right)Gg- \\
\left(F_2-G\right)\left(f_1-g\right)Gg+\left(f_1-g\right)\left(f_2-g\right)G^2= \\
\left(F_1-G\right)\left(F_2-G\right)g^2+\left(f_1-g\right)\left(f_2-g\right)G^2-\left(F_1+f_1-G-g\right)\left(F_2+f_2-G-g\right)gG+ \\
\left(F_1-G\right)\left(F_2-G\right)gG+\left(f_1-g\right)\left(f_2-g\right)gG= \\
\left(F_1-G\right)\left(F_2-G\right)g\left(G+g\right)+\left(f_1-g\right)\left(f_2-g\right)G\left(G+g\right)- \\
\left(F_1+f_1-G-g\right)\left(F_2+f_2-G-g\right)gG.
\end{gathered} \tag{13}$$

Using (13) and (12) we have

$$\begin{gathered}
P\left(Y_1\wedge Y_2>Y_3\vee Y_4\,|\Theta\right)+P\left(Y_1\wedge Y_2<Y_3\vee Y_4\,|\Theta\right)- \\
P\left(Y_1\wedge Y_3>Y_2\vee Y_4\,|\Theta\right)-P\left(Y_1\wedge Y_3>Y_2\vee Y_4\,|\Theta\right)= \\
\sum\left(F_1-G\right)\left(F_2-G\right)g+\sum\left(F_1-G\right)\left(F_2-G\right)g\left(1-G-g\right)+ \\
\sum\left(F_1+f_1-G-g\right)\left(F_2+f_2-G-g\right)g+\sum\left(F_1+f_1-G-g\right)\left(F_2+f_2-G-g\right)gG+ \\
\sum\left(f_1-g\right)\left(f_2-g\right)G\left(1-G-g\right)+\int\left(F_1-G\right)\left(F_2-G\right)\tilde{g}\,\mathrm{dy}\,.
\end{gathered}$$

We therefore conclude that conditionally on $\left\{X_1=x^{(1)},X_2=x^{(2)}\right\}$

$$\begin{gathered}
P\left(Y_1\wedge Y_2>Y_3\vee Y_4,X_3\vee X_4<X_1\wedge X_2\right)+ \\
P\left(Y_1\wedge Y_2<Y_3\vee Y_4,X_3\vee X_4<X_1\wedge X_2\right)- \\
P\left(Y_1\wedge Y_3>Y_2\vee Y_4,X_3\vee X_4<X_1\wedge X_2\right)- \\
P\left(Y_1\wedge Y_3<Y_2\vee Y_4,X_3\vee X_4<X_1\wedge X_2\right)= \\
\sum\left(P\left(Y<y\,|\,X=x^{(1)}\right)-P\left(Y<y\right)\right)\left(P\left(Y<y\,|\,X=x^{(2)}\right)-P\left(Y<y\right)\right)\cdot \\
P\left(Y=y\,\middle|\,X<x^{(1)}\wedge x^{(2)}\right)\left(P\left(X<x^{(1)}\wedge x^{(2)}\right)\right)^2+ \\
\sum\left(P\left(Y<y\,|\,X=x^{(1)}\right)-P\left(Y<y\right)\right)\left(P\left(Y<y\,|\,X=x^{(2)}\right)-P\left(Y<y\right)\right)\cdot \\
P\left(Y=y\,\middle|\,X<x^{(1)}\wedge x^{(2)}\right)P\left(Y>y\,\middle|\,X<x^{(1)}\wedge x^{(2)}\right)\left(P\left(X<x^{(1)}\wedge x^{(2)}\right)\right)^2+ \\
\sum\left(P\left(Y\le y\,|\,X=x^{(1)}\right)-P\left(Y\le y\right)\right)\left(P\left(Y\le y\,|\,X=x^{(2)}\right)-P\left(Y\le y\right)\right)\cdot \\
P\left(Y=y\,\middle|\,X<x^{(1)}\wedge x^{(2)}\right)\left(P\left(X<x^{(1)}\wedge x^{(2)}\right)\right)^2+ \\
\sum\left(P\left(Y\le y\,|\,X=x^{(1)}\right)-P\left(Y\le y\right)\right)\left(P\left(Y\le y\,|\,X=x^{(2)}\right)-P\left(Y\le y\right)\right)\cdot
\end{gathered}$$

$$P\left(Y=y\,\middle|\,X<x^{(1)}\wedge x^{(2)}\right)P\left(Y<y\,\middle|\,X<x^{(1)}\wedge x^{(2)}\right)\left(P\left(X<x^{(1)}\wedge x^{(2)}\right)\right)^2+$$
$$\sum\left(P\left(Y=y\mid X=x^{(1)}\right)-P\left(Y\leq y\right)\right)\left(P\left(Y=y\mid X=x^{(2)}\right)-P\left(Y=y\right)\right)\cdot$$
$$P\left(Y<y\,\middle|\,X<x^{(1)}\wedge x^{(2)}\right)P\left(Y>y\,\middle|\,X<x^{(1)}\wedge x^{(2)}\right)\left(P\left(X<x^{(1)}\wedge x^{(2)}\right)\right)^2+$$
$$\int\left(P\left(Y<y\mid X=x^{(1)}\right)-P\left(Y<y\right)\right)\left(P\left(Y<y\mid X=x^{(2)}\right)-P\left(Y<y\right)\right)\cdot$$
$$P\left(Y\in \mathrm{dy}\,\middle|\,X<x^{(1)}\wedge x^{(2)}\right)\left(P\left(X<x^{(1)}\wedge x^{(2)}\right)\right)^2.$$

All of the above terms lead to non-negative expressions because of Lemma 2 (for the first, third and sixth term we take $C=\Omega$, the set of all possible outcomes). We also see that these terms can be zero iff X and Y are independent. □

References

Agresti, A. (2010). *Analysis of ordinal categorical data (second edition).* New York: Wiley.

Bergsma, W. P. (2006). A new correlation coefficient, its orthogonal decomposition, and associated tests of independence. *arXiv:math/0604627v1 [math.ST].*

Blum, J. R., Kiefer, J., & Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *The annals of mathematical statistics*, *32*, 485-498.

Dassios, A., & Bergsma, W. (2012). Supplementary material for: "a consistent test of independence based on a sign covariance related to kendall's tau". *Technical report.*

De Wet, T. (1980). Cramér-von Mises tests for independence. *J. Multivariate Anal.*, *10*, 38-50.

Deheuvels, P. (1981). An asymptotic decomposition for multivariate distribution-free tests of independence. *J. Multivariate Anal.*, *11*, 102-113.

Feuerverger, A. (1993). A consistent test for bivariate dependence. *International Statistical Review / Revue Internationale de Statistique*, *61*(3), pp. 419-433.

Hoeffding, W. (1948). A non-parametric test of independence. *Annals of Mathematical Statistics*, *19*, 546-557.

Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods.* NY: Wiley.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), pp. 81-93.

Kendall, M. G., & Gibbons, J. D. (1990). *Rank correlation methods.* New York: Oxford University Press.

Kimeldorf, G., & Sampson, A. R. (1978). Monotone dependence. *Ann. Stat.*, *6*(4), pp. 895-903.

Kruskal, W. H. (1958). Ordinal measures of association. *J.Am.Stat.Ass*, *53*, 814-861.

Nelsen, R. B. (2006). *An introduction to copulas.* New York: Springer.

Rényi, A. (1959). On measures of dependence. *Acta Math. Acad. Sci. Hung.*, *10*, pp. 441-451.

Schweizer, B., & Wolff, E. F. (1981). On nonparametric measures of dependence for random variables. *Annals of Statistics*, *9*, 879-885.

Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures (fourth edition).* Boca Raton: Chapman and Hall.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), pp. 72-101.

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Stat.*, *35*(6), 2769-2794.

Wilding, G. E., & Mudholkar, G. S. (2008). Empirical approximations for Hoeffding's test of bivariate independence using two Weibull extensions. *Stat. Methodol.*, *5*(2), 160–170.