On a new measure of rank-order association

Agostino Tarsitano (*)
Dipartimento di Economia, Statistica e Finanza - Università della Calabria
Via Pietro Bucci, Cubo 1c, 87036 Rende (CS) - Italy
agostino.tarsitano@unical.it

Ilaria L. Amerise
Dipartimento di Economia, Statistica e Finanza - Università della Calabria
Via Pietro Bucci, Cubo 1c, 87036 Rende (CS) - Italy
ilaria.amerise@unical.it

Abstract:

Rank correlations currently in use have a resistance-to-change which appears to be of limited value for the purposes of ranking comparisons. It is plain that a given value of a rank correlation does not define a specific pair of permutations, except perhaps for the extreme values. Nevertheless, a coefficient that condenses comparison of rankings into too few values renders difficult the assessment of the strength of their association. Recently, Tarsitano & Lombardo [2013] proposed a new statistic of rank correlation, called r_4, based on the intuitive appeal of quotients, which achieves greater sensitivity to changes in rankings than any other known coefficient; and this without causing additional difficulty in interpretation or affecting the implementation in hypothesis testing. In there, the authors do not discuss the finite and limiting behavior of the new coefficient. In the present paper we show that the exact null distribution of r_4 is well approximated by the t-Student ad that, its asymptotic distribution, is a standard Gaussian distribution. Computational results for empirical and simulated data sets reveal that r_4 is very efficient in evaluating strength and pattern of the agreement between pairs of rankings.

Keywords

Asymptotic Gaussianity, Independence tests, Ordinal association, Rank statistics.

(*) Corresponding author.

1. Introduction

Dependence between rankings is a topic that persistently occurs throughout statistical practice and it is the subject of the present paper. Our point of departure is the fact that, though the Pearson's product-moment correlation coefficient (here denoted r_0) is widely used to measure the linear relation between two variables, it can perform poorly when the relationship is thought to be non-linear and/or the data are affected by errors of measurement and outliers. For example, it needs only one abnormal value to shift r_0 to any value in the interval $[-1,1]$. For these and many other reasons, we may turn to more resistant, albeit less efficient non-parametric measure of association.

Consider n independent pairs of scores $(x_i,y_i), i=1,2,\cdots,n$. The pairs are sorted into ascending order of their first coordinate and then transformed into the ranks $\pi = \{\pi_1,\pi_2,\cdots,\pi_n\}$. Likewise, the y_i, $i=1,2,\cdots,n$ are placed in correspondence with the ranks $\eta = \{\eta_1,\eta_2,\cdots,\eta_n\}$. Both π and η are elements of S_n, the set of all $n!$ permutations of the integers $\{1,2,\cdots,n\}$. With no essential loss of generality we assume that π_i is the rank of x_i after η has been arranged in its natural order, that is $\eta_i = i, i=1,2,\cdots,n$. Note, also, that we assume there are no ties throughout.

A rank correlation $r(\eta,\pi)$ is a statistic summarizing the degree of agreement between η and π. Three of the more popular rank correlation coefficients are:

$$\begin{aligned}
&\textit{Spearman} \quad r_1(\pi,\eta) = \frac{12}{n^3-n}\sum_{i=1}^{n} i\pi_i - 3\left(\frac{n+1}{n-1}\right)\\
&\textit{Kendall} \quad r_2(\pi,\eta) = \frac{2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\operatorname{sgn}(\pi_j-\pi_i)}{n(n-1)}\\
&\textit{Gini} \quad r_3(\pi,\eta) = \frac{4\left[\sum_{n+1-i\le\pi_i}[\pi_i-(n+1-i)]-\sum_{i\le\pi_i}(\pi_i-i)\right]}{n^2-k_n}
\end{aligned} \tag{1}$$

with $k_n = n \mod 2$ and $\operatorname{sgn}(.)$ equals to $-1,0$ or 1 according to whether its argument is negative, zero or positive. We note that r_1 takes values on a lattice with $\left(n^3-n\right)/6+1$ distinct values. The sum $\sum_{i=1}^{n} i\pi_i$ in the first term of r_1 covers all the integers between

$n(n+1)(n+2)/6$ and $n(n+1)(2n+1)/6$. When $n>3$, r_1 can be zero if, and only if, n is not of the form $n = 4*m+2$ where m is a positive integer (see Marshall, 1994). The possible values of r_2 are $\left(n^2-n\right)/2+1$. The coefficient is zero or even if, and only if, $n = 4*m$ or $n = m*4+1$ where m is any positive integer; r_2 only takes on odd values if n is not in that form. When $n>3$, zero is always a value of Gini's coefficient r_3, which can assume other $2\left(n^2/4+k_n\right)$ distinct values. In each case, the expression within square brackets in r_3 only takes on even values. According to Kendall [1938], the disparity in the potential number of values among rank correlations is not a great disadvantage to their sensitivity. Nonetheless, Kendall & Gibbons [1990][p. 37-38] used this argument to dismiss Spearman's footrule as a feasible measure of association.

The choice of a rank correlation is fundamentally based on two antithetical requirements: resistance and sensitivity. Resistance refers to the ability of a coefficient to remain constant when data are changed slightly. However, since stability is achieved at the cost of a loss in precision, it can become a problem if the same value is applied to describe very different patterns. Sensitive coefficients offer a richer source of information regarding the association structure, but sensitivity is a drawback when substantially similar rankings are mapped onto distant coefficient values. A reasonable compromise may be achieved by considering that, since ranks rely on the relative ordering of elements, they are, by construction, very tolerant of noise and disturbances that do not affect the actual order. Thus, particular consideration should be given to the discriminatory power of a coefficient rather than to its resistance. From this point of view, many robust rank correlations such those proposed by Dallal & Hartigan [1980], Blomqvist [1950]) or Gideon & Hollister [1987] are largely insufficient for ranking comparisons when the range of possible relationships between the underlying variables is wide.

Apparently, coefficients in (1) have the right characteristics to act as valid substitutes of Pearson's correlation whenever it is necessary. Nonetheless, the spectrum of their values is still relatively small and concentrated on a reduced set of points. Tarsitano

& Lombardo [2013], proposed a new rank correlation coefficient based on the intuitive appeal of quotients

$$r_4(\pi,\eta)=\frac{(\mathbf{b}_{\eta,\pi^*})^t\,\mathbf{b}_{\eta^*,\pi}-(\mathbf{b}_{\eta^*,\pi^*})^t\,\mathbf{b}_{\eta,\pi}}{M_n},\quad M_n=\left[k_n+2\sum_{i=1}^{\lfloor n/2\rfloor}(n+1-i)/i\right]^2-n^2. \quad (2)$$

where $\lfloor x\rfloor$ denotes the largest integer not greater than x. The symbols $\pi^*=n+1-\pi$ and $\eta^*=n+1-\eta$ are the reverse permutations of π and η, respectively. The $n\times 1$ vector $\mathbf{b}_{\eta,\pi}$ is formed with the components of the matrix $\mathbf{A}$ occupying the positions identified by the elements in η as first index and those in π as second index. The generic element of $\mathbf{A}$ is $a_{ij}=\max(i,j)/\min(i,j), i,j=1,2,\cdots,n$. The coefficient r_4 can assume a number of distinct values of the order $0.25n!$ more or less uniformly spaced from each other.

Coefficients $r_h(\eta,\pi), h=1,\cdots,4$ share several properties, notably monotonicity, symmetry, right-invariance and antisymmetry under reversal. See Gideon & Hollister [1987] and Brown & Eagleson [1984]. All the coefficients vary within the range: $[-1,1]$. The extremes are achieved if and only if there is perfect association for all pairs: $r_h(\eta,\eta)=r_h(\pi,\pi)=1$, $r_h(\eta,\eta^*)=r_h(\pi,\pi^*)=-1$. The closer r_h (for brevity, the arguments π,η are dropped unless ambiguity occurs) is to one, ignoring the sign, the stronger the relationship between rankings is. At the other extreme, $r_h=0$ or near-zero implies that the two rankings are not related according to the association concept embodied in r_h.

In Figure 1 the exact null distributions of $r_1,\cdots,r_4$ are shown as frequency polygons for $n=10$. The profiles show some resemblances to and some differences from one another. The frequency polygons of r_2 and r_3 exhibit high levels of irregularities. We attribute this to the lattice of values available for these coefficients, which is much sparser than that of r_1 or r_4. In fact, the space between possible values of r_2 and r_3 decreases monotonically, but slowly as n increases. A good sign, however, is that the serration is more noticeable in the middle of the range $[-1,1]$ than near the extremes where they have a greater importance for hypothesis testing. The varying size of serrations in the frequency polygon of r_1 is less intense than those in r_2 and r_3, but much more sharp than that of r_4. The profile of r_4 shows the slightest degree of fluctuation and the tails of its null distribution smooth out first and more than any of the other coefficients.

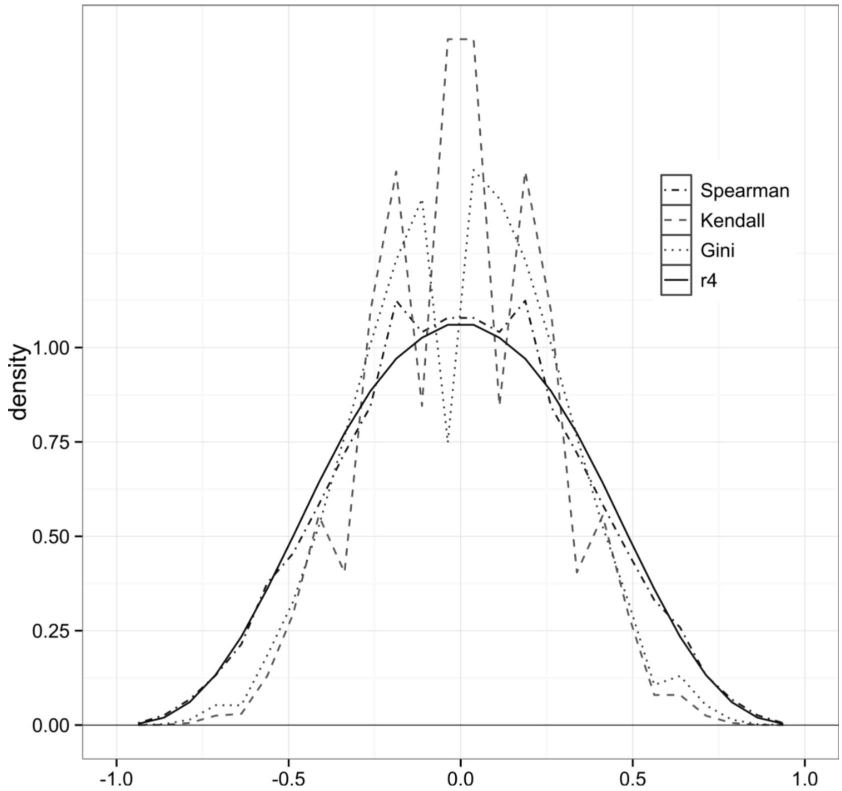

Figure 1: Frequency polygons (based on binned counts) for n=10.

2. Sampling distribution of r_4 under independence

In this section, we are concerned with the distribution of r_4 when all rankings are equally probable with probability $1/n!$. Firstly, the denominator of r_4 is unaffected by any permutation of the ranks so that it is sufficient to consider the random variable $M_n r_4 = \mathbf{b}^t_{\eta,\pi^*}\mathbf{b}_{\eta^*,\pi} - \mathbf{b}^t_{\eta^*,\pi^*}\mathbf{b}_{\eta,\pi}$, which has support in $[-M_n, M_n]$. The properties of r_4 ensure that, for each pair of permutations such that $\mathbf{b}^t_{\eta,\pi^*}\mathbf{b}_{\eta^*,\pi} = x$, there must be another pair of permutations for which also $\mathbf{b}^t_{\eta^*,\pi^*}\mathbf{b}_{\eta,\pi} = x$ and, consequently, $\mathbf{b}^t_{\eta,\pi^*}\mathbf{b}_{\eta^*,\pi}$ and $\mathbf{b}^t_{\eta^*,\pi^*}\mathbf{b}_{\eta,\pi}$ share the same codomain. If follows that $E\left(\mathbf{b}^t_{\eta,\pi^*}\mathbf{b}_{\eta^*,\pi}\right) = E\left(\mathbf{b}^t_{\eta^*,\pi^*}\mathbf{b}_{\eta,\pi}\right)$ which, in turn, implies that $E(r_4) = 0$. Hence, under the null hypothesis of independent rankings, the distribution of r_4 is symmetrical around zero and has support in $[-1,1]$. All the odd moments are zero because of the symmetry.

The calculation of the variance is more difficult than that of the mean. We have

$$\begin{aligned} M_n^2 V(r_4) &= V\left(\mathbf{b}_{\eta,\pi^*}^t \mathbf{b}_{\eta^*,\pi} - \mathbf{b}_{\eta^*,\pi^*}^t \mathbf{b}_{\eta,\pi}\right) \\ &= V\left(\mathbf{b}_{\eta,\pi^*}^t \mathbf{b}_{\eta^*,\pi}\right) + V\left(\mathbf{b}_{\eta^*,\pi^*}^t \mathbf{b}_{\eta,\pi}\right) - 2Cov\left(\mathbf{b}_{\eta,\pi^*}^t \mathbf{b}_{\eta^*,\pi}, \mathbf{b}_{\eta^*,\pi^*}^t \mathbf{b}_{\eta,\pi}\right). \end{aligned} \tag{3}$$

By virtue of the same reasoning as used above for the derivation of the expected value, we obtain $V\left(\mathbf{b}_{\eta,\pi^*}^t \mathbf{b}_{\eta^*,\pi}\right) = V\left(\mathbf{b}_{\eta^*,\pi^*}^t \mathbf{b}_{\eta,\pi}\right)$. Then, expression (3) specializes to

$$V(r_4) = \frac{2\left[V\left(\mathbf{b}_{\eta,\pi^*}^t \mathbf{b}_{\eta^*,\pi}\right) - Cov\left(\mathbf{b}_{\eta,\pi^*}^t \mathbf{b}_{\eta^*,\pi}, \mathbf{b}_{\eta^*,\pi^*}^t \mathbf{b}_{\eta,\pi}\right)\right]}{M_n^2}. \tag{4}$$

We have empirically explored (4) by evaluating it over all possible pairs of permutations with n up to 15 and found that, under independence, the Pearson correlation coefficient $cor\left(\mathbf{b}_{\eta,\pi^*}^t \mathbf{b}_{\eta^*,\pi}, \mathbf{b}_{\eta^*,\pi^*}^t \mathbf{b}_{\eta,\pi}\right)$ converges towards -1 as n increases. Based on this premise, (4) can be reasonably approximated by

$$\sigma_n^2(r_4) \approx \frac{4\left[V\left(\mathbf{b}_{\eta,\pi^*}^t \mathbf{b}_{\eta^*,\pi}\right)\right]}{M_n^2}. \tag{5}$$

It remains to evaluate the variance of the dot-product $V\left(\mathbf{b}_{\eta,\pi^*}^t \mathbf{b}_{\eta^*,\pi}\right)$. One limitation of our paper is that we were not able to write (5) in a simplified manner, even exploiting the relationships developed by Bohrnstedt & Goldberger [1969] and Brown & Eagleson [1984] on the exact variance and covariance of a product of random variables. To circumvent this problem, we apply a simply linear regression model

$$\sigma_n^2(r_4) = \frac{\beta}{n-1} + \varepsilon. \tag{6}$$

The regression function has no intercept to allow the variance to reach zero as n goes to infinity. In passing we note that (6) coincides with the asymptotic variance of Spearman's coefficient r_1 when $\beta = 1$. The true values of $\sigma_n^2(r_4)$ are determined by complete enumeration of all rankings. The unknown parameter β is estimated by the linear least squares method applied to the 11 points $\left[\sigma^2(r_4), n\right], n = 5, \cdots, 15$. The resulting estimate is $\sigma_n^2(r_4) \approx 1.00762/(n-1)$ with an adjusted R^2 of 0.9994. This approximation is quite good even for small values of n as it is shown in the first two rows of Table 1.

Table 1: Exact and approximate values of $\sigma_n^2(r_4)$.

n	7	8	9	10	11	12	13	14	15
$\sigma_n^2(r_4)$	0.1677	0.1423	0.1275	0.1131	0.1037	0.0945	0.0879	0.0815	0.0766
$\hat{\sigma}_n^2(r_4)$	0.1679	0.1439	0.1260	0.1120	0.1008	0.0916	0.0840	0.0775	0.0720
$\sigma_n^2(r_1)$	0.1667	0.1429	0.1250	0.1111	0.1000	0.0909	0.0833	0.0769	0.0714
$\sigma_n^2(r_2)$	0.1005	0.0833	0.0710	0.0617	0.0545	0.0488	0.0442	0.0403	0.0370
$\sigma_n^2(r_3)$	0.1204	0.0982	0.0875	0.0756	0.0689	0.0614	0.0569	0.0518	0.0485

In the last three rows Table 1 we report the variances of the Spearman, Kendall and Gini coefficients, which show that the distribution of r_4 is relatively more disperse than that of the other rank correlations.

The coefficient of kurtosis of r_4 can also be obtained through the same regression strategy. The corresponding least squares estimate is

$$\hat{\gamma}_n(r_4) \approx 2.929894 - \frac{5.889006}{n} + \frac{8.559322}{n^2} - \frac{11.617287}{n^3} \tag{7}$$

with an adjusted R^2 virtually equal to one and a residual standard error of 0.000116. Thus, $\gamma_n(r_4)$ converges to a limit value near three (the value of kurtosis for a Gaussian distribution) as n goes to infinity. We show in Table 2 the results of the fitting procedure. The interpolation of $\gamma_n(r_4)$ is excellent and would be quite satisfactory in practice. This result is particularly important in the present work, since there does not appear to be any simple way in which either moments or cumulants of r_4 can be determined. The last rows in Table 2 reports the kurtosis values of r_1, r_2 and r_3, which confirm that r_4 is slightly more platykurtic than the other coefficients

Table 2: Exact and approximate values of $\gamma_n(r_4)$.

n	7	8	9	10	11	12	13	14	15
$\gamma_n(r_4)$	2.2292	2.3049	2.3653	2.4150	2.4565	2.4918	2.5222	2.5487	2.5719
$\hat{\gamma}_n(r_4)$	2.2294	2.3048	2.3653	2.4150	2.4565	2.4919	2.5223	2.5487	2.5719
$\gamma_n(r_1)$	2.3357	2.4190	2.4840	2.5360	2.5785	2.6140	2.6440	2.6696	2.6919
$\gamma_n(r_2)$	2.6833	2.7262	2.7586	2.7839	2.8043	2.8211	2.8351	2.8471	2.8574
$\gamma_n(r_3)$	2.5238	2.5310	2.6213	2.6078	2.6869	2.6615	2.7335	2.7007	2.7682

2.1. *t*-Student approximation

In consideration of the affinities between r_4 and r_1, at least for the first three moments, we suggest a procedure similar to that used by Zar [1972] and Landenna *et al.* [1989]. Let r be a random variable with a Pearson type II density

$$f(r,\lambda) = \frac{\left(1-r^2\right)^{(\lambda-1)}}{B(0.5,\lambda)} \qquad \text{with } |r| \leq 1; \quad \lambda > 0. \tag{8}$$

where B is the well-known beta function and λ is a parameter positively related to the number of ranks n. The variance and kurtosis of r are

$$\sigma^2(\lambda) = 1/(2\lambda+1), \qquad \gamma(\lambda) = -6/(2\lambda+3) \tag{9}$$

The variance decreases monotonically as λ, and hence n, increases. The kurtosis is negative denoting that (8) is less peaked and has thinner tails than the Gaussian distribution. For $\lambda \to \infty$, the Pearson type II density becomes quite close to the standardized Gaussian density. See Devroye [1986][p. 433]. On the other hand, for $\lambda \to 0^+$, the general lower bound on symmetrical densities: $\gamma(\lambda) > -2$ is verified. See Devroye [1986][p. 688]. In summary, curve (8) is symmetrical, unimodal with mode at zero, supported within interval $[-1,1]$ and has a tendency towards the Gaussian distribution. If we set

$\sigma^2(r) = 1.00762/(n-1)$ and solve the first equation in (9) for λ then r has approximately the same variance as r_4 and a kurtosis roughly equal to $\gamma(r) = -6.04572/(n+2.01524)$. The difference between the kurtosis of r_4 and that of r becomes negligible as $n \to \infty$.

One key factor behind the wide diffusion of (8) is its strict relationship with the Student's t density function, which allows for the use of easy tables and hence ensures computational convenience and simple checking of results. In particular, the following statistic $r_4' = r_4\sqrt{2m/\left(1-r_4^2\right)} \sim t_{\lfloor 2m \rfloor}$ with $m = (n-1.00762)/2.01524$ can be used to test the significance of r_4. See, for example, Willink [2009]. The quality of the approximations is illustrated in Table 3. For the given α, we report the exact conservative critical value, the approximated critical value and their absolute difference.

Table 3: Comparison of t-Student approximation to the exact distribution of r_4.

n	α	Exact	Approx.	Abs. Dif.	α	Exact	Approx.	Abs. Dif.
12	0.0001	0.8815	0.8947	0.0133	0.0100	0.6647	0.6851	0.0204
	0.0005	0.8295	0.8470	0.0175	0.0250	0.5833	0.6021	0.0188
	0.0010	0.8009	0.8199	0.0190	0.0500	0.5053	0.5214	0.0161
	0.0025	0.7556	0.7759	0.0203	0.1000	0.4065	0.4187	0.0122
	0.0050	0.7142	0.7348	0.0206	0.2500	0.2229	0.2281	0.0052
13	0.0001	0.8649	0.8742	0.0093	0.0100	0.6448	0.6581	0.0132
	0.0005	0.8111	0.8233	0.0122	0.0250	0.5643	0.5760	0.0117
	0.0010	0.7819	0.7950	0.0130	0.0500	0.4877	0.4973	0.0096
	0.0025	0.7360	0.7496	0.0136	0.1000	0.3913	0.3981	0.0067
	0.0050	0.6943	0.7079	0.0136	0.2500	0.2138	0.2161	0.0023
14	0.0001	0.8451	0.8544	0.0092	0.0100	0.6239	0.6339	0.0100
	0.0005	0.7902	0.8010	0.0108	0.0250	0.5446	0.5529	0.0083
	0.0010	0.7607	0.7717	0.0110	0.0500	0.4697	0.4762	0.0065
	0.0025	0.7145	0.7255	0.0110	0.1000	0.3761	0.3802	0.0041
	0.0050	0.6729	0.6835	0.0107	0.2500	0.2048	0.2058	0.0009
15	0.0001	0.8302	0.8353	0.0051	0.0100	0.6076	0.6120	0.0045
	0.0005	0.7743	0.7800	0.0057	0.0250	0.5293	0.5324	0.0032
	0.0010	0.7444	0.7501	0.0058	0.0500	0.4557	0.4575	0.0018
	0.0025	0.6979	0.7034	0.0055	0.1000	0.3642	0.3646	0.0003
	0.0050	0.6563	0.6614	0.0051	0.2500	0.1979	0.1968	0.0010

From Table 3 it can be seen that the accuracy of approximation tends to be lower for smaller α. When n increases, the general quality of approximation improves and turns to be higher where it is most needed, that is, in the tails of the distribution.

3. Large-sample distribution of r_4

In case n is too large for complete enumeration to be feasible, the distribution of r_4 can be approximate by using a continuous curve such as the t-Student density. If, however, there is no special reason (other than a good fit) to use a particular probability density, we can resort to the Gaussian density and rely on some form of the central limit theorem.

Let us define $\zeta\left(\eta_i, \eta_j, \pi_i, \pi_j\right) = g\left(\eta_i, \pi_i^*\right) g\left(n+1-\eta_j, \pi_j\right) - g\left(n+1-\eta_i, \pi_i^*\right) g\left(\eta_j, \pi_j\right)$. The quantity $g_i(\pi, \eta) = \exp\{|\log(\pi_i) - \log(\eta_i)|\}$, $i = 1, 2, \cdots, n$ expresses the disagreement between two rankings due to the distance from π_i to η_i. By construction, $E(G_n) = 0$. It is important to notice that G_n clearly falls within the class of double-indexed permutation statistics studied by Zhao *et al.* [1997] (see also Barbour & Chen, 2005). The crucial result, for our purposes, is Theorem 2 in Zhao *et al.* [1997] in which the authors, by using the Stein's method, prove that there is a constant $K > 0$ such that for $n \geq 2$

$$\sup_x \left| P\left(G_n \leq \sigma(G_n) x\right) - \Phi(x) \right| \leq \frac{K}{\sigma(G_n)^3} \left\{ n^{-1} \sum_{i,k} \left| a_{i,k}^* \right|^3 + \sum_{i,j,k,l} \left| \zeta_{i,j,k,l}^* \right|^3 \right\} \tag{10}$$

where $\Phi(x)$ is the standard Gaussian distribution and

$$\begin{aligned} a_{i,k} &= \zeta_{i,i,k,k}^* + n^{-1} \sum_{j,l} \zeta_{i,j,k,l} + n^{-1} \sum_{j,l} \zeta_{j,i,l,k} \\ a_{i,k}^* &= a_{i,k} - \sum_{k=1}^{n} a_{i,k} - \sum_{i=1}^{n} a_{i,k} + \sum_{k=1}^{n} \sum_{i=1}^{n} a_{i,k} \end{aligned} \tag{11}$$

with

$$\begin{aligned} \zeta_{i,j,k,l}^* = \zeta_{i,j,k,l} - n^{-1} &\left[\sum_{l} \zeta_{i,j,k,l} + \sum_{k} \zeta_{i,j,k,l} + \sum_{j} \zeta_{i,j,k,l} + \sum_{i} \zeta_{i,j,k,l} \right] + \\ + n^{-2} &\left[\sum_{k,l} \zeta_{i,j,k,l} + \sum_{j,l} \zeta_{i,j,k,l} + \sum_{j,k} \zeta_{i,j,k,l} + \sum_{i,l} \zeta_{i,j,k,l} + \sum_{i,k} \zeta_{i,j,k,l} + \sum_{i,j} \zeta_{i,j,k,l} \right] \\ &- n^{-3} \left[\sum_{k,j,l} \zeta_{i,j,k,l} + \sum_{i,k,l} \zeta_{i,j,k,l} + \sum_{i,j,l} \zeta_{i,j,k,l} + \sum_{i,k,j} \zeta_{i,j,k,l} \right]. \end{aligned} \tag{12}$$

The condition to be satisfied for the validity of (10) is $M_n\sigma^2(G_n) = \sum_{k=1}^{n}\sum_{i=1}^{n}\left(a^*_{i,k}\right)^2 > 0$. This is simply an estimate of the variance of G_n, which, as we have argued, can be asymptotically approximated by $\sigma^2(G_n) \approx M_n(n-1)^{-1}$. Applying (10), we can conclude that the null distribution of $r^*_4 = r_4/\sigma_n(r_4)$ converges to $\Phi(x)$ with the rate $O(1/\sqrt{n})$.

The point that we want to emphasize is that the large-sample approximation to the exact null distribution of r_4, suitably standardized, can be based on the Gaussian distribution. For this standardization, it is necessary to know expected value and variance of r_4 when the hypothesis of independence is true. We have shown in the previous section that, under such hypothesis, $E(r_4) = 0$ and $\sigma^2(r_4) \approx 1.00762(n-1)^{-1}$. It follows that $r^*_4 = 1.003803 r_4\sqrt{n-1}$ has, for n tending to infinity, an asymptotic Gaussian distribution.

To demonstrate the applicability of the limiting distribution to the null, we investigate r_4 together with Spearman's r_1. This coefficient is taken as benchmark reference because it is very widely known, but above all, because an important aim of our article is to understand whether there is any evidence that a large number of potential values give an advantage to the discriminatory power of a rank correlation. In this sense, the variety of values of r_1 is the richest among the statistics commonly in use at the present time.

From Figure 2 we see that, while the agreement between the frequency polygon of r_4 and the Gaussian curve is not adequate in the middle, it is satisfactory in the wings *i.e.* precisely where it is more necessary for testing independence. However, since the frequency polygon of r_4 is shorter in the tails than the corresponding Gaussian curve, using this as an approximation can lead to a test that is more liberal than necessary, that is, the null hypothesis of independence will tend to be rejected more frequently than it should be.

Further insights can be gained by Table 4 in which the proportions of total frequencies falling outside the ranges $[-a, a]$ for $a = 1, 1.25, 2, 2.5, 3$ predicted by the Gaussian model

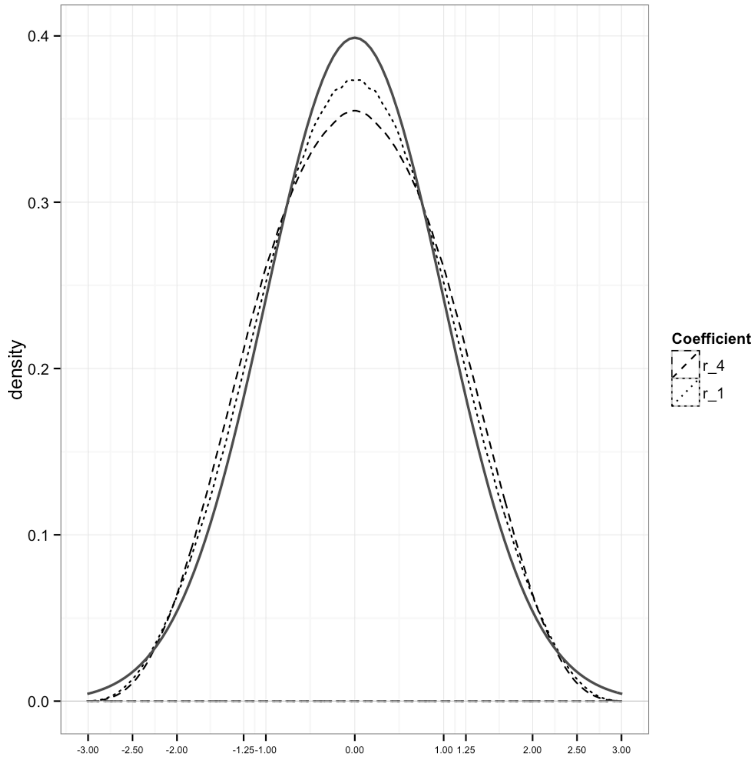

Figure 2: Comparison of the Gaussian approximation (thick line) with the exact null distribution of r_4 (dashed lines) andr_1 (dotted line) for $n = 12$.

are compared with those observed in the exact null distribution of r_4 and r_1.

Table 4: Proportion of frequencies of the distribution of r_4 and r_1 falling in certain ranges.

n	Coefficient	$\pm\sigma$	$\pm 1.25\sigma$	$\pm 2\sigma$	$\pm 2.5\sigma$	$\pm 3\sigma$
	Gaussian	0.6827	0.8944	0.9545	0.9876	0.9973
11	r_4	0.6419	0.7589	0.9598	0.9955	1.0000
	r_1	0.6585	0.7750	0.9598	0.9945	1.0000
12	r_4	0.6440	0.7599	0.9583	0.9946	0.9999
	r_1	0.6690	0.7724	0.9601	0.9938	0.9999
13	r_4	0.6423	0.7574	0.9555	0.9933	0.9998
	r_1	0.6658	0.7760	0.9598	0.9933	0.9997
14	r_4	0.6431	0.7575	0.9542	0.9925	0.9997
	r_1	0.6668	0.7790	0.9581	0.9928	0.9996
15	r_4	0.6415	0.7553	0.9519	0.9914	0.9995
	r_1	0.6665	0.7788	0.9578	0.9924	0.9995

The Gaussian density yields liberal results especially for high values (in absolute terms) of the transformed rank correlations and it is conservative within intervals roughly from ± -0.75 to ± -2.25. The frequency polygons of r_4^* and r_1^* deviate quite considerably from Gaussianity in the interval $[-1.25, 1.25]$ implying that significance levels at around 20 percent are largely overestimated. The approximation is acceptably accurate for significance levels barely above 5%, but fails, although not so spectacularly, for smaller levels.

4. Experimental results

In the preceding sections, we have discussed the exact null distribution of r_4, the new rank correlation proposed by Tarsitano & Lombardo [2013], as well as the Gaussian and t-Student approximations. The aim of the present section is to provide a guide to the correct use of r_4 in empirical research and to highlight some potential misuse through applying it to real and simulated data sets.

The algorithms described in this section are implemented in the a package *pvrank* in the *R* system (R Development Core Team [2013]), which is available from the authors on request.

4.1. Real data examples

We have selected four data sets that are briefly described below. For each of these, we provide the scatter plot with a vertical and horizontal line drawn at the mean values of the variables. In addition, we create a summary table of the test: $H_0: r_h = 0$ against the two-sided alternative $H_0: r_h \neq 0, h = 0, 1, \cdots, 4$. It should be recalled that, as correctly observe Iman & Conover [1978], the discreteness of rank correlations often leads into situations where no critical region has exactly the size α. Rather there will be a choice of using the next smaller exact size called conservative p-value (denoted by C_α) or the next larger exact size called liberal p-value (L_α). Clearly, this consideration does not apply when the null distribution is approximated by a continuous distribution.

Example 1. Hollander & Wolfe [1999][p. 39]. Hamilton depression scale factor measurements in $n = 9$ patients with mixed anxiety and depression, taken at the first and second visit after initiation of a therapy. See graph a) in Figure 3. Apparently, there are no outliers, so that rank correlations and significance levels should not fall too far from the values obtained for r_0. The results in Table 5 confirm that this is the case for r_4 and only partially for r_1. What is more serious still is that the p-values associated with r_2 are doubtful at $\alpha = 0.05$ (those of r_3 are doubtful at $\alpha = 0.10$).

Example 2. In this case, we use the data set *CWD* (Hothorn *et al.*, 2013). An infrared gas analyzer and a clear chamber sealed to the wood surface were used to measure the flux of carbon out of the wood. Measurements were repeated $n = 13$ times. Although not necessarily linear, there is a general decrease in Y as X increases. See graph b). The findings reported in Table 5 send contradictory signals as to the association strength. Specifically, r_1, r_2 and r_3 suggest that there is a more significant relationship between the ranks of X and Y than what is suggested by r_4. On the other hand, coefficient r_4 gives the most similar results to those of Pearson's r_0.

Example 3. Here, we consider the data set in Berk [1990] including data on the average number of births and deaths by the time of the day for a particular hospital in Brussels. We have discarded pairs in which at least one element is repeated and remained with $n = 19$ valid data points. As evident from the graphs c) and d), seventeen observations are clustered and show little association. Two observations (for noon and midnight) are dramatically smaller in both the y-direction and x-direction. With these two included, there is obviously a positive correlation in the data. However, a direct association between the two variables is questionable, for if the outliers are removed then all correlations decrease and the associated p-values increase up to the point where the hypothesis of independence cannot be rejected at any reasonable level. Note that, r_4 achieves, in both testing situations, the nearest proximity to r_0 while conserving a good degree of robustness against the effect of outliers. Furthermore, when the outliers are removed, r_4 has the

Table 5: Measure of correlation/association and p-values

Index	symbol	obs.	C_α	L_α	obs.	C_α	L_α
		Hamilton data			*CWD data*		
Pearson	r_0	0.8479	0.00388	0.00388	-0.5256	0.06507	0.06507
Spearman	r_1	0.6500	0.06656	0.07604	-0.6484	0.01816	0.01941
Kendall	r_2	0.5000	0.04462	0.07518	-0.4872	0.01495	0.02158
Gini	r_3	0.5250	0.07447	0.11079	-0.5238	0.02062	0.02801
	r_4	0.6918	0.03923	0.03925	-0.5782	0.04335	0.04335
		Births and deaths by the hour			*Outliers removed*		
Pearson	r_0	0.6802	0.00135	0.00135	0.1556	0.55100	0.55100
Spearman	r_1	0.3877	0.10190	0.10356	0.1446	0.57886	0.58544
Kendall	r_2	0.2865	0.08007	0.09330	0.1029	0.54233	0.59764
Gini	r_3	0.2778	0.14596	0.16301	0.0694	0.71715	0.76700
	r_4	0.5319	0.02026	0.02026	0.2269	0.38020	0.38020
		Urban percentage			*Outlier removed*		
Pearson	r_0	-0.6212	0.01774	0.01774	-0.7882	0.00137	0.00137
Spearman	r_1	-0.5385	0.04786	0.04996	-0.7418	0.00461	0.00508
Kendall	r_2	-0.3846	0.04718	0.06166	-0.5385	0.00668	0.01012
Gini	r_3	-0.4898	0.02438	0.03174	-0.6190	0.00475	0.00716
	r_4	-0.5231	0.06191	0.06191	-0.7207	0.00654	0.00654

lowest (albeit non significant) p-value among all the statistics based on ranks, which is an indicator of its sensitivity to changes in rankings.

Example 4. This example is taken from Birkes & Dodge [1993]. The data set report birth rate and urban percentage for $n = 14$ countries in North and Central America. The data point 13 (corresponding to Trinidad-Tobago) stands far apart from the rest of the points. The possible effect on the measures of correlation and association is a low value of the statistics even if there is a very apparent association between variables. In fact, once the outlier is excluded from the data set, the p-values of all the coefficients decrease of a factor of ten. Actually, if the outlier is included, only coefficient r_4 is not significantly different from zero (at the 5% level or lower), whereas the p-values of the other statistics seem scarcely affected by the outlier. Rather than a defect, we consider this low resistance to the impact of outliers as a virtue that adds flexibility to the use of r_4.

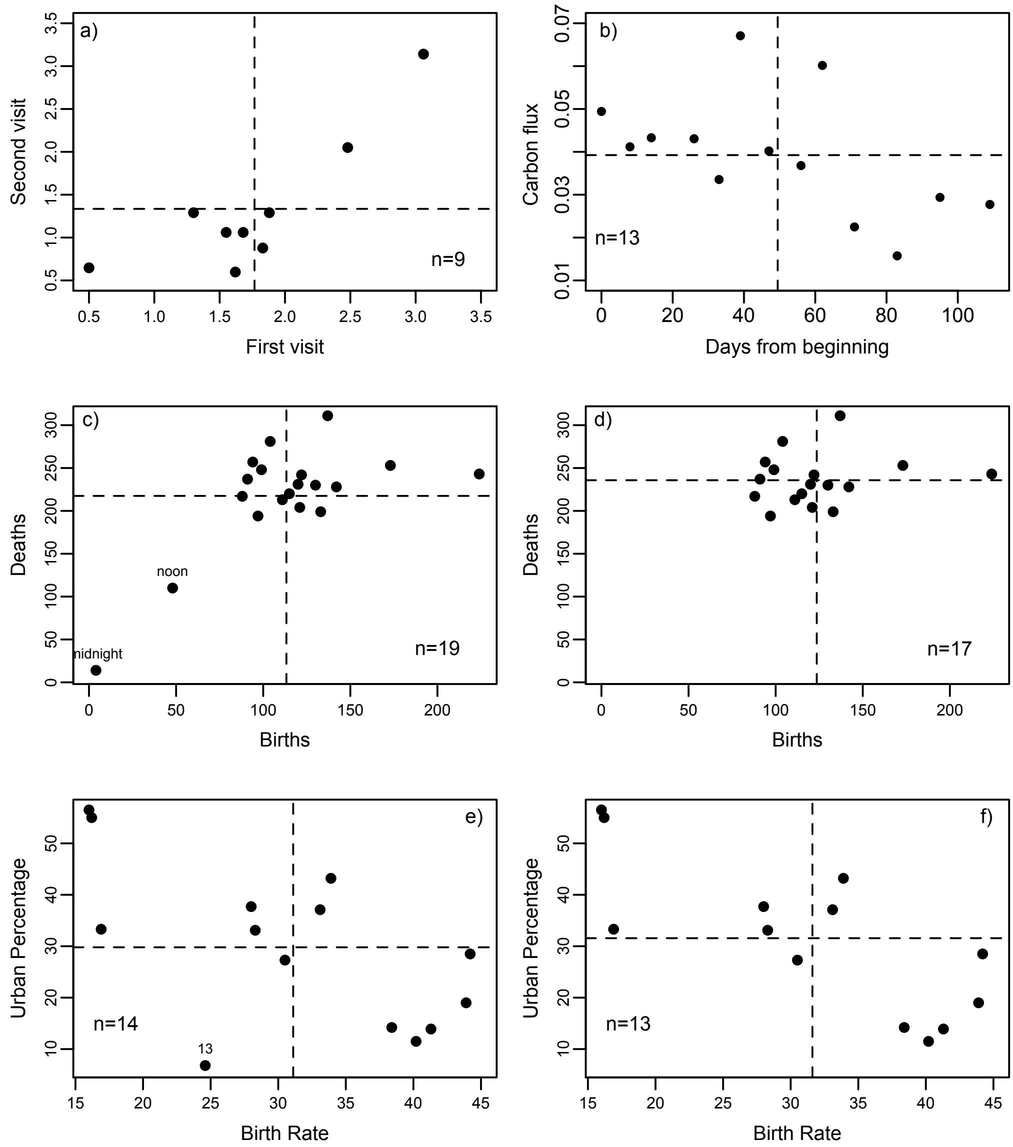

Figure 3: Type of association discussed in the examples.

The findings in Table 5 suggest that r_4 (based on ranks) is an admissible substitute for r_0, (based on scores). A useful feature of r_4 is that, because of its high resolution over the set of all permutations, conservative and liberal p-values almost coincide and, therefore, the risk of doubtful testing is reduced with respect to the other three rank correlations. Furthermore, the richness of the range of values renders its intrinsic discrete nature so marginal that the effect of a continuity correction, either beneficial or detrimental, is negligible.

4.2. Simulation

To assess the power performance of the test corresponding to r_4 we carried out the following experiments. First, we generate independent samples $(x_i, y_i), i = 1, \cdots, n$ of size $n = 10$ and $n = 15$ from bivariate Gaussian populations with means zero, variances of one and zero correlation. The generation is repeated until $N = 10,000$ samples are formed.

To avoid occasionally significant correlation, we have excluded samples with an r_0 outside $[-0.20, 0.20]$. Second, k outliers are introduced. Let $(i_1, \cdots, i_h)$ be the set of integers from $1, \cdots, n$ such that $x_{i_1} y_{i_1} > 0, \cdots, x_{i_h} y_{i_h} > 0$. If $h < k$, then the sample is discarded. The pairs $\left(x_{i_j}, y_{i_j}\right), j = 1, \cdots, h$ are sorted into descending order of their Euclidean distance from the origin. The first k pairs of observations are contaminated by displacing their values by m standard deviations in both the x- and y-direction. This induces spurious positive correlation that tends to increase with the numbers of outliers and the amount of displacement. In Table 6 we compare the numbers of samples declared significant at the alpha level (one-tail) of 1%, 5% and 10% by using the t-student distribution with $(n-2)$ degrees of freedom in the case of r_0 and the exact null distributions for r_h, $h = 1, \cdots, 4$.

The number of rejections of $H_0 : r_0 = 0$ against $H_1 : r_0 > 0$ at level α is greater with $n = 10$ than with $n = 15$. This result is to be expected because the exceptional nature of some observations is more perceivable when the same numbers of outliers occur in a wider and otherwise homogeneous sample. In addition, the numbers of samples producing a false positive correlation increase with the magnitude of the shift. Even this result is not surprising given that a large displacement makes the artificial outliers manifestly inconsistent with the regression model (a line parallel to the x axis) that is called on implicitly. Furthermore, in line with the expectations, the numbers of wrong claims become more severe with a greater numbers of abnormal data points.

The behavior described above is also exhibited by rank correlations, but with one fundamental difference: the number of wrong rejections is now much less than with Pearson's correlation. In this regard, we observe a different behavior for mild contamination, *i.e.* $m = 1,2$, and wider contamination, *i.e.* $m = 3,4$. In the former case, the statistic that has the smallest number of improper rejections is most often r_3. In the latter case is r_2. Perhaps it is helpful to note that r_2 and r_3 have the smallest range of possible values.

Table 6: Number of significant samples (over 10,000) for $r_0, \cdots, r_4$.

			Pearson r_0 α level			Spearman r_1 α level			Kendall r_2 α level			Gini r_3 α level			r_4 α level		
n	k	m	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10
10	1	1	0	11	263	0	1	7	0	4	9	0	4	10	0	3	22
		2	148	2096	3843	0	1	13	0	7	13	0	4	13	0	6	50
		3	1921	5074	6353	0	3	21	0	7	16	0	8	20	0	8	69
		4	4243	6604	7410	0	5	29	0	10	28	0	11	24	0	12	91
	2	1	0	56	691	0	2	28	0	10	37	0	6	22	0	27	139
		2	346	2749	4626	0	8	93	1	27	74	0	14	48	1	71	401
		3	2036	5014	6479	0	16	151	1	42	113	0	27	71	1	118	635
		4	3491	6052	7242	0	27	199	2	55	138	0	39	89	6	159	789
	3	1	0	106	914	0	14	112	0	32	75	0	17	45	2	85	427
		2	355	2800	4883	2	63	405	7	96	243	0	52	134	8	372	1509
		3	1746	4858	6513	3	118	687	8	152	409	1	89	223	12	660	2394
		4	2849	5711	7104	4	178	897	8	212	538	1	121	296	19	892	3021
15	1	1	0	153	1006	0	3	17	0	7	33	0	9	49	0	8	55
		2	1006	5402	8600	0	4	26	0	11	50	0	11	61	0	18	114
		3	5617	11035	13162	0	6	36	0	12	58	0	16	75	0	27	159
		4	9918	13720	15167	0	9	45	0	16	73	0	20	82	0	32	195
	2	1	2	636	2367	0	6	60	0	19	95	0	18	87	1	72	442
		2	1921	7097	10609	0	13	166	2	43	187	1	35	159	2	212	1094
		3	5976	11500	14152	0	22	257	2	62	268	1	50	214	2	335	1605
		4	8761	13428	15512	0	37	325	3	79	314	1	65	259	7	433	1943
	3	1	7	951	3170	0	24	210	0	51	211	0	35	171	5	285	1215
		2	2124	7600	11484	2	97	683	8	146	572	1	84	382	22	1053	3576
		3	5611	11518	14482	3	184	1097	9	257	871	2	153	561	41	1747	5347
		4	7806	13119	15555	4	269	1408	10	336	1108	2	203	698	60	2265	6577

In general, the figures in Table 6 simply reaffirm, what is already well known, that the Pearson correlation coefficient can produce incorrect indications if outliers affect data. More importantly, in the presence of anomalies, rank correlations are more reliable in the assessment of evidence of a relationship between two variables. The coefficient r_4 occupies an intermediate position between Pearson's product-moment correlation and the standard statistics of rank-order association: Spearman, Kendall and Gini. On one hand, r_4 may sporadically produce erroneous, significant associations as it is shown by the slightly inflated alpha level for the some combinations k and m. On the other hand, it is capable of capturing even weak relationship between the variables otherwise lost in case other measures of association are applied.

References

Barbour, A. D. and Chen, L. H. Y. (2005). "The permutation distribution matrix correlation statistics" In: Barbour, A. D.; Chen, L. H. Y. *Stein's method and applications*, 223 – 245. Singapore University Press.

Berk, R. A. (1990). "A primer on robust regression". In Fox, J. and Scott Long, J. (Eds), 292 – 324. *Modern Methods of Data Analysis*. Sage Publications, Newbury Park, Ca, USA.

Birkes, D. and Dodge, Y. (1993). *Alternative Methods of Regression*. John Wiley & Sons, New York.

Blomqvist, N. (1950). "On a measure of dependence between two random variables". *The Annals of Mathematical Statistics*, **21**, 593 – 600.

Bohrnstedt, G. W. and Goldberger, A. S. (1969). "On the exact covariance of products of random variables" *Journal of the American Statistical Association*, **64**,1439 – 1442.

Brown, T. C. and Eagleson G. K. (1984)."A useful property of some symmetric statistics". *The American Statistician*, **38**, 63 – 65.

Dallal, G. E. and Hartigan, J. A. (1980). "Note on a test of monotone association insensitive to outliers". *Journal of the American Statistical Association*, **75**, 722 – 725.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.

Gideon, R. A. and Hollister, A. (1987). "A rank correlation coefficient resistant to outliers. *Journal of the American Statistical Association*, **82**, 656 – -666.

Hothorn, T. and Hornik, K. and van de Wiel M. A. and Zeileis, A. (2013). coin: Conditional inference procedures in a permutation test framework. http://CRAN.R-project.org/package=coin. R package version 1.0-23.

Hollander, M. and Wolfe, D. A. (1999). *Nonparametric Statistical Methods. 2nd edn*. John Wiley & Sons, New York.

Iman, L. and Conover, W. J. (1978). "Approximations of the critical region for Spearman's rho with and without ties present. *Communication in Statistics - Simulation and Computation* **7**, 269 – 282.

Kendall, M. G. (1938). "A new measure of rank correlation" *Biometrika*, **30**, 81 – 93.

Kendall, M. G. and Gibbons, J. D. (1990). *Rank Correlation Methods, 5th edn*. Oxford University Press. New York.

Landenna, G. and Scagni, A. and Boldrini, M. (1989),"An approximated distribution of the Gini's rank association coefficient" *Communications in Statistics. Theory and Methods*, **18**, 2017 – 2026.

Marshall, E. I. (1994). "Conditions for rank correlation to be zero" *Sankhya: The Indian Journal of Statistics, Series B*, **56**, 59 – 66.

R Core Team *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 2013; available at http://www.R-project.org/.

Tarsitano, A, Lombardo, R. (2013). "A coefficient of correlation based on ratios of ranks and anti-ranks". *Jahrnbucher für Nationalökonomie und Statistik* **233**, 206 – 224.

Willink R. A. (2009). Single form for t-distributions and symmetric beta distributions. *Communications in Statistics - Theory and Methods* **39**, 170 – 176.

Zar, J. H. (1972). Significance testing of the spearman rank correlation coefficient. *Journal of the American Statistical Association*, **67**, 578 – 580.

Zhao, L. and Bai, Z. and Chao, C. -C. and Liang, W.-Q. (1997). "Error bound in a central limit theorem of double-indexed permutation statistics" *The Annals of Statistics*, **25**, 2210 – 2227.