Journal of Statistics: Advances in Theory and Applications
Volume 21, Number 1, 2019, Pages 23-51
Available at http://scientificadvances.co.in
DOI: http://dx.doi.org/10.18642/jsata_7100122031

AN ALTERNATIVE TEST OF RANK-ORDER ASSOCIATION IN THE PRESENCE OF TIES

ILARIA L. AMERISE

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

Abstract

A novel non-parametric test of rank-order association, denoted as r_4, has been recently introduced to address the problem of finding a robust coefficient that, at the same time, is highly sensitive to differences between rankings. Coefficient r_4 has advantages over some best-known rank-correlations such as Spearman's ρ, Kendall's τ, and Gini's γ, in respect of the smoothness of the sampling distribution and fineness of the resolution.

The properties of r_4 have been established for cases in which rankings do not contain ties. It is the purpose of the present paper to deal with cases in which ties are present. Several treatments of ties need to be considered to find the most appropriate one. By using real and simulated data sets, we find that the Gideon-Hollister method gives the best procedure to establish how strongly or weakly two variables are associated when at least one of the rankings contains ties.

2010 Mathematics Subject Classification: 62GXX, 62H10, 62G30.
Keywords and phrases: ordinal data, robust rank correlation, independence testing, non-parametric statistics.
Received February 6, 2019

© 2019 Scientific Advances Publishers

1. Introduction

Pearson's product-moment correlation coefficient, denoted with r_0, is widely used as a summary measure of the strength of the relationship between two variables. However, it is well known that it can perform poorly when 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 $[-1, 1]$. In view of this lack of robustness, it is simpler to think in terms of non-parametric coefficients of association that are more resistant, albeit less efficient from the point of view of the adherence to the values.

Let (x_i, y_i), $i = 1, 2, \cdots, n$ be a sample n independent pairs drawn from a bivariate population with joint distribution $H(X, Y)$ and marginal distributions $F(X)$ and $G(Y)$. An important component of many statistical analyses is a measure of association between X and Y. A reasonably robust strategy would be to express association in terms of ranks. The use of ranks considerably simplifies mathematical analysis by placing the intervals between successive ranked values on a common scale.

Let us sort the pairs (x_i, y_i) into ascending order of their first coordinate and then transform them into the ranks $\eta = \{\eta_1, \eta_2, \cdots, \eta_n\}$. Normally, ranks are integer numbers $(1, 2, 3, \cdots)$. Likewise, the second coordinates y_i, $i = 1, 2, \cdots, n$ are placed in correspondence of the first coordinates with the ranks $\pi = \{\pi_1, \pi_2, \cdots, \pi_n\}$. Both π and η are elements of S_n, the set of all $n!$ permutations of the integers $\{1, 2, \cdots, n\}$.

A rank correlation $r\{\eta, \pi\}$ is intended to measure to which extent a monotonic function between the two rankings is able to model the inherent relationship between the two variables once the original observations have been transformed into ranks. Let $\rho(\eta, \pi)$ be the rank

correlation that would be obtained averaging $r(\eta, \pi)$ over the entire set S_n of all the permutations of the integers $1, 2, \cdots, n$. Hoeffding [11] established that, if $n \geq 5$, then the test of hypothesis

$$\begin{cases} H_0 : H(X, Y) = F(X)G(Y) \\ H_1 : H(X, Y) \neq F(X)G(Y) \end{cases} \text{ is equivalent to } \begin{cases} H_0 : \rho(\eta, \pi) = 0 \\ H_1 : \rho(\eta, \pi) \neq 0. \end{cases} \quad (1)$$

Rank correlations neither assume a specific parametric model nor specific distributions of the variables. In short, the test on the left can be performed by evaluating the hypothesis that each ranking $\pi = \{\pi_1, \pi_2, \cdots, \pi_n\}$ has the same probability $(1 / n!)$ of being associated to η. Large deviations from zero are evidence against independence.

There are many coefficients of rank-correlation. Obviously, we restrict our attention to coefficients that have an expected value of zero under H_0. In spite of this restriction, coefficients ranges widely. The choice of the most suitable one is based on two antithetical requirements. First, the ability of a coefficient to remain constant when experimental conditions are changed slightly. Second, sensitivity to changes in rankings. Since stability is achieved at the cost of a loss in precision, it may become a problem if the same value is applied to describe very different situations. 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 good compromise between robustness and sensitivity is offered by coefficient r_4 (Tarsitano & Lombardo [19], Tarsitano & Amerise [20]) which not only exploits the intuitive appeal of quotients of ranks, but also have a high resolution over the set of all permutations. In those papers, it is agued that three well known rank correlation measures: Spearman's ρ, Kendall's τ, and Gini's γ are not ideally suited for measuring rank

correlation for numerical data that are perturbed by noise. Consequently, a more elastic rank correlation measure, such as r_4, seems promising with respect to the expression and differentiation of the relationship between variables.

The properties of r_4, however, have been reviewed only in those instances where observed values or preference lists do no contain ties. Therefore, the influence of ties on this test has not been investigated. This is a shortcoming of the procedure, because in real problems we almost always have cases when two observations are exactly equal due to rounding or two or more alternatives are objectively equivalent by the given criterion.

It is necessary to premise that equal ranks should be seen as a disease to be cured. When ties are present, the permutations are not all distinct, the number of possible permutations is reduced and the range of rank correlations becomes narrower than the conventional margins $[-1, 1]$. Thus, new margins must be determined for each specific configuration of tied observations and for each coefficient. Experimental research workers must be aware that a proportion of tied scores, which is relatively large in comparison with n, alters the null distribution of all rank-correlation, so caution should be used in evaluating the significance level of the tests. However, even when the number of ties is small, an inefficient method of treating them may lead to situations in which a rank correlation cannot yield meaningful results. In summary, researchers should make every effort to avoid equal values for different observations.

Our objective in the present paper, is to assess the behaviour of r_4 when, despite best attempts, at last one of the rankings contain ties. The structure of the paper is as follows. In the next section, we review the properties of r_4 in comparison with those of the most commonly used rank-correlations. In particular, we discuss the exact distribution of the r_4 coefficient under the hypothesis of independence. Additionally, the

large-sample distribution theory for r_4 is described. Section 3 explores the behaviour of r_4 in the event that some of the rankings contain ties. The final section summarizes the paper contributions and presents some conclusions.

2. Properties of r_4

In this section, coefficient r_4 is re-examined in comparison with three popular measures that are often provided next to each other by standard software packages, namely Spearman's rank-order correlation, Kendall's tau coefficient, and Gini's cograduation index. The expressions of such coefficients are

$$\begin{aligned}
&\textit{Spearman} \quad r_1(\pi) = \frac{12}{n^3 - n}\sum_{i=1}^{n}\eta_i\pi_i - 3\left(\frac{n+1}{n-1}\right),\\
&\textit{Kendall} \quad r_2(\pi) = \frac{2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\operatorname{sgn}(\pi_j - \pi_i)}{n(n-1)},\\
&\textit{Gini} \quad r_3(\pi) = \frac{4\left[\sum_{\eta_i^* \le \pi_i}(\pi_i - \eta_i^*) - \sum_{i \le \pi_i}(\pi_i - \eta_i)\right]}{n^2 - k_n},\\
&\textit{Tarsitano} \quad r_4(\pi) = \frac{\det A_n(\pi)}{M_n},
\end{aligned} \tag{2}$$

where $\eta_i = i, 1 = 1, \cdots, n$ is the identity permutation, $\eta_i^* = n + 1 - i$, $i = 1, 2, \cdots, n$ are the reverse or complementary ranks, k_n equals to 0,1 according to whether n is even or odd and sgn(x) equals to – 1, 0, 1 according to whether x is negative, zero or positive. The denominator of r_4 is

$$M_n = \frac{1}{n^2}\left\{\left[k_n + 2\sum_{i=1}^{\lfloor n/2 \rfloor}(n + 1 - i)/i\right]^2 - n^2\right\}, \tag{3}$$

where $\lfloor x \rfloor$ denotes the largest integer not greater than x. Furthermore, the numerator of r_4 is the determinant of the following (2×2) matrix

$$\det A_n(\pi) = \frac{1}{n} \begin{pmatrix} a_\pi = \left[\sum_{i=1}^{n} \left\langle \frac{\eta_i}{\pi_i^*} \right\rangle \right] & b_\pi = \left[\sum_{i=1}^{n} \left\langle \frac{\eta_i}{\pi_i} \right\rangle \right] \\ c_\pi = \left[\sum_{i=1}^{n} \left\langle \frac{\eta_i^*}{\pi_i^*} \right\rangle \right] & d_\pi = \left[\sum_{i=1}^{n} \left\langle \frac{\eta_i^*}{\pi_i} \right\rangle \right] \end{pmatrix} = a_\pi d\pi - b_\pi c_\pi, \tag{4}$$

where $\pi_i^* = n + 1 - \pi_i$, $i = 1, 2, \cdots, n$ is the reverse ranks of π. The symbol $\langle . \rangle$ appearing in (4) denotes the maximum ratio between two positive numbers

$$\left\langle \frac{x_i}{y_i} \right\rangle = \max\left\{ \frac{x_i}{y_i}, \frac{y_i}{x_i} \right\}, \quad \text{for } x_i, y_i > 0, \; i = 1, \cdots, n. \tag{5}$$

Then each element in the matrix $A_n(\pi)$ will have n quotients. See Mango [14] and Zhang [22]. Notice that M_n is the maximum of the absolute difference between $a_\pi d_\pi$ and $b_\pi c_\pi$.

Coefficients $r_h(\pi)$, $h = 1, \cdots, 4$ share several properties, notably monotonicity, symmetry, right-invariance and antisymmetry under reversal (see Gideon & Hollister [8]). All the coefficients vary within the range: $[-1, 1]$. The extremes are achieved if and only if there is perfect association for all pairs: $r_h(\pi, \pi) = 1$, $r_h(\pi, \pi^*) = -1$. The closer r_h (for brevity, the π argument is 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 some values of n.

A relevant question is whether a value of r_4 reflects a genuine measure of agreement/disagreement between two rankings, or whether the value has just happened by chance. We are thus led to examine the null distribution of r_4, that is the distribution of r_4 when a given ranking is compared with all $n!$ permutation in S_n. It must be pointed out, that the null distribution of rank correlations depends only on the sample size n. The only difficult is the calculation of the observed statistic, while the null distribution can be computed once for all and critical values stored to be applied to any dataset of size n.

In this spirit, the exact null distribution of $r(\eta, \pi)$ is fully available for many rank correlation measures and for values of n and their applicability tends to increase as computing resources are extended further and further. See Girone et al. [10] (Gini), Panneton & Robillard [17] (Kendall), Maciak [13] (Spearman). Package pvrank (Amerise et al., [2]) contains the exact distribution of coefficients r_1 (Spearman), r_2 (Kendall), r_3 (Gini) for $n = 26, 60, 24$, respectively.

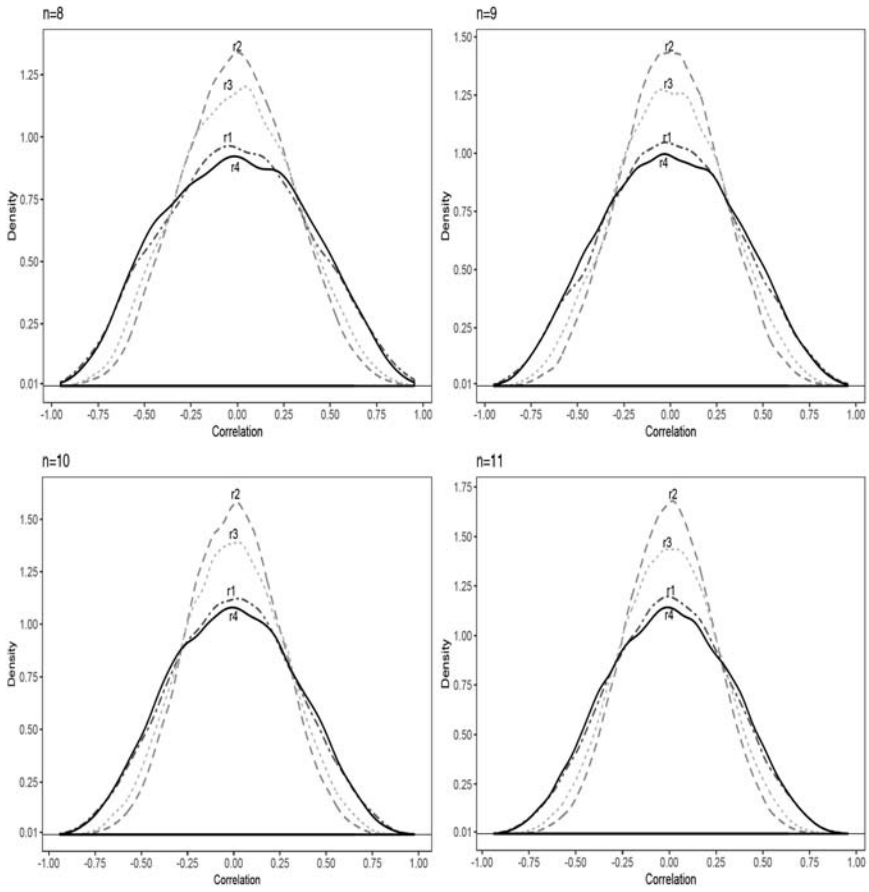

Figure 1. Frequency polygons (based on binned counts) for n = 8, 9, 10, 11.

2.1. Exact sampling distribution of r_4

Firstly, we notice that, for given n, the denominator of r_4 is fixed and does not depend on the permutation π, so that it is sufficient to look at

$$M_n r_4 = (a_\pi d_\pi - b_\pi c_\pi). \tag{6}$$

The properties of r_4 ensure that, for each permutations such that $a_\pi d_\pi = x$, there is another pair of permutations for which $b_\pi c_\pi = x$ also

and, consequently, a_π and b_π share the same codomain. It follows that $E(a_\pi) = E(b_\pi)$ which, in turn, implies that $E(r_4) = 0$. Hence, under the hypothesis of independent rankings, the distributions of r_4 are symmetrical around zero and have support in $[-1, 1]$. All the odd moments are zero because of the symmetry. By virtue of the same reasoning as used above, it is wholly clear that

$$E(a_\pi) = E(b_\pi) = E(c_\pi) = E(d_\pi),\ \sigma^2(a_\pi) = \sigma^2(b_\pi) = \sigma^2(c_\pi) = \sigma^2(d_\pi). \quad (7)$$

Analogously, we have $\sigma^2(a_\pi d_\pi) = \sigma^2(b_\pi c_\pi)$. It follows that

$$\begin{aligned} M_n^2\sigma_n^2(r_4) &= \sigma^2(a_\pi d_\pi) + \sigma^2(b_\pi c_\pi) - 2Cov(a_\pi d_\pi,\ b_\pi c_\pi) \\ &= \sigma^2(a_\pi d_\pi) + V(b_\pi c_\pi) - 2Cor(a_\pi d_\pi,\ b_\pi c_\pi)\sigma(a_\pi d_\pi)\sigma(b_\pi c_\pi) \\ &= 2\sigma^2(a_\pi d_\pi) - 2Cor(a_\pi d_\pi,\ b_\pi c_\pi)\sigma^2(a_\pi d_\pi). \end{aligned} \quad (8)$$

Therefore, the variance of r_4 specifies to

$$\sigma^2(r_4) = \frac{2\sigma^2(a_\pi d_\pi)[1 - Cor(a_\pi d_\pi,\ b_\pi c_\pi)]}{M_n^2}. \quad (9)$$

We have empirically explored (9) by evaluating it over all possible pairs of permutations, with n up to 15 and found that, under independence, the product-moment correlation coefficient $Cor(a_\pi d_\pi,\ b_\pi c_\pi)$ converges rapidly towards -1 as n increases. Based on this premise, (9) can be fairly approximated by

$$\sigma_n^2(r_4) = \frac{4\sigma^2(a_\pi d_\pi)}{M_n^2}. \quad (10)$$

Unfortunately, we have not been able to bring the variance corresponding to r_4 into a form suitable for numerical computation. For this reason, the linear regression model

$$\sigma_n^2(r_4) = \frac{\beta}{n-1} + \varepsilon, \tag{11}$$

is used to fit (by least squares) the data on $\sigma_n^2(r_4)$ against n. The regression function has no intercept to allow the variance to reach zero as n goes to infinity. 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 12 points $[\sigma^2(r_4), n]$, $n = 4, \cdots, 15$. 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 satisfactory even for small n as it is shown in the first two rows of Table 1. In the last three rows, we report the variances of the Spearman, Kendall and Gini coefficients. The estimates are indicated with a hat. It is most evident that the estimated variances are very close to the actual ones and both indicate that the distribution of r_4 is relatively more disperse than that of the other rank correlations.

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

n	7	8	9	10	11	12	13	14	15
$\sigma^2(r_4)$	0.1677	0.1423	0.1275	0.1131	0.1037	0.0945	0.0879	0.0815	0.0766
$\widehat{\sigma}_n^2(r_4)$	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

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

$$\hat{\gamma}_n(r_4) \approx 2.929894 - \frac{5.889006}{n} + \frac{8.559322}{n^2} - \frac{11.617287}{n^3}, \qquad (12)$$

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 the results of the fitting procedure in Table 2. The interpolation of $\gamma(r_4)$ 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 three 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
$\gamma(r_4)$	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

2.2. Limiting distribution of r_4

In a situation where exact results are not available, it is natural to derive asymptotic distributions relying on some form of the central limit theorem (see Cifarelli et al., [6]). Hopefully, these will give accurate

approximations that can be used in establishing approximate critical regions for testing purposes. As preliminary step in this direction, we note that, in general, the logarithm of the maximum ratio between two positive numbers is given by

$$\log\left(\left\langle\frac{x}{y}\right\rangle\right) = |\log(x) - \log(y)|. \tag{13}$$

To simplify the development that follows, let us define a quantity depending on n

$$\begin{aligned}\zeta_n(\eta_i,\, \eta_j,\, \pi_i,\, \pi_j) &= g(\eta_i,\, n+1-\pi_i)g(n+1-\eta_j,\, \pi_j) \\ &\quad - g(n+1-\eta_i,\, n+1-\pi)g(\eta_j,\, \pi_j),\end{aligned} \tag{14}$$

where

$$g(\pi_i,\, \eta_j) = g(\pi_j,\, \eta_i) = \exp\{|\log(\pi_i) - \log(\eta_j)|\},\ i,\, j = 1,\, 2,\, \cdots,\, n, \tag{15}$$

expresses the disagreement between two rankings due to the distance between π_i and η_j. Recalling that $\eta_i = i$, $i = 1, 2, \cdots, n$, the numerator of r_4 becomes

$$M_n r_4 = G_n = \sum_{i,\, j=1}^{n} \zeta_n(i,\, j,\, \pi_i,\, \pi_j). \tag{16}$$

By construction, $E(G_n) = 0$. The statistic G_n falls within the class of double-indexed permutation statistics studied by Zhao et al. [21] (see also Barbour & Chen [4]). The crucial result, for our purposes, is Theorem 2 in Zhao et al. [21] in which the authors, by using the Stein's method, prove that there is a constant $K > 0$ such that for $n \geq 2$, the following inequality holds

$$\sup_x |P[G_n \leq x\sigma(G_n)] - \Phi(x)| \leq \frac{K}{\sigma(G_n)^3}\left\{n^{-1}\sum_{i,k}\left|a^*_{i,k}\right|^3 + \sum_{i,j,k,l}\left|\zeta^*_{i,j,k,l}\right|^3\right\}, \tag{17}$$

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

$$a_{i,k} = \zeta^*_{i,i,k,k} + n^{-1}\sum_{j,l}\zeta_{i,j,k,l} + n^{-1}\sum_{j,l}\zeta_{j,i,l,k},$$

$$a^*_{i,k} = a_{i,k} - \sum_{k=1}^{n} a_{i,k} - \sum_{i=1}^{n} a_{i,k} + \sum_{k=1}^{n}\sum_{i=1}^{n} a_{i,k}, \tag{18}$$

with

$$\zeta^*_{i,j,k,l} = \zeta_{i,j,k,l} - n^{-1}\left[\sum_{l}\zeta_{i,j,k,l} + \sum_{k}\zeta_{i,j,k,l} + \sum_{j}\zeta_{i,j,k,l} + \sum_{i}\zeta_{i,j,k,l}\right]$$

$$+ n^{-2}\left[\sum_{k,l}\zeta_{i,j,k,l} + \sum_{j,l}\zeta_{i,j,k,l} + \sum_{j,k}\zeta_{i,j,k,l} + \sum_{i,l}\zeta_{i,j,k,l} + \sum_{i,k}\zeta_{i,j,k,l} + \sum_{i,j}\zeta_{i,j,k,l}\right]$$

$$- n^{-3}\left[\sum_{k,j,l}\zeta_{i,j,k,l} + \sum_{i,k,l}\zeta_{i,j,k,l} + \sum_{i,j,l}\zeta_{i,j,k,l} + \sum_{i,k,j}\zeta_{i,j,k,l}\right]. \tag{19}$$

The condition to be satisfied for the validity of (17) is

$$\sigma^2(G_n) = \sum_{k=1}^{n}\sum_{i=1}^{n}(a^*_{i,k})^2 > 0. \tag{20}$$

The variance of G_n is proportional to the variance of r_4, that is, $\sigma^2(G_n) = M_n^2\sigma^2(r_4)$. Therefore, we have $\sigma^2(G_n) \approx M_n^2 1.00762(n-1)^{-1}$. It is simply, but tedious to show that M_n^2 converges to 1729.640125 so that condition (20) is fulfilled.

By applying (17), 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, may be based on the Gaussian distribution. We showed 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 an asymptotic Gaussian distribution for n tending to infinity.

To illustrate that the limiting distribution can be applied to the null, we investigate r_4 together with Spearman's r_1. In Table 3, 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 are compared with those observed in the exact null distribution of r_4 and r_1. Except at the end of the scale, the Gaussian distribution considerably underestimates the probability.

Table 3. Proportion of frequencies of the distribution of r_4 and r_1 falling in certain ranges

n	Coefficient	$\pm\sigma$	$\pm1.25\sigma$	$\pm2\sigma$	$\pm2.5\sigma$	$\pm3\sigma$
	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 approximation yields liberal thresholds 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 agreement with the distribution of r_4 is far from satisfactory, though it is not sensibly worse than that between the Gaussian and the distribution of Spearman r_1. The frequency polygons of the standardize versions of r_4 and r_1 deviate quite considerably from Gaussianity in the [–1.25, 1.25] interval implying that significance levels at around 20 percent are largely overestimated. The approximation is acceptably accurate for significance levels that are barely above 5%, but fails, although not spectacularly so, for smaller levels.

3. The Treatment of Ties in the r_4 Test

The independence test $H_0 : H(X, Y) = F(X)G(Y)$ has been developed for continuous random variables. When F and G are continuous, the probability of getting tied observations is zero, so that this event may be ignored. However, the condition of continuity is not appropriate when data consist of integer-valued or the attributes are measured with limited precision or data are censored or measured only on an ordinal scale. In such settings equal values appear frequently in one or both rankings. In this section we describe how we can extend r_4 to work with ties.

We begin by assuming that the values $\{x_1, x_2, \cdots, x_n\}$ are set in ascending order. If the same score is repeated for more than one observation, we arrange the values into k_x mutually exclusive and collectively exhaustive groups where each group contains $t_{x,i}$ equal values occupying the positions from $j_{x,i,1}$ to $j_{x,i,t_{x,i}}$ with $\sum_{i=1}^{k_x} t_{x,i} = n$. There are m_x groups with $t_i > 1$. The $n - m_x$ groups for which $t_{x,i} = 1$

(which we define as fixed positions) receive the integer corresponding to their positions in the ranking from 1 to n. The values in the other m_x groups are represented either by the set of ranks they would have had if they were not tied or by a set formed by repeating $t_{x,i}$ times the mean (not necessarily the arithmetic mean) of the ranks pertaining that block. The values $\{y_1, x_y, \cdots, y_n\}$ undergo the same process of classification in k_y blocks, m_y of which include ties with $\sum_{i=1}^{k_y} t_{y,i} = n$, independently of what happens in the other ranking. The ranks of untied observations are well defined but those of tied observations are not. When ties are present, the value of r_4 is not uniquely determined, and it is necessary to adopt some supplementary modifications of its definition.

Ties can be broken in numerous ways yielding different values of r_4 ranging from, say, r_4^- to r_4^+ and will possibly lead to different conclusions. The symbol r_4^- denotes the value of r_4 for a pair of permutations in which ties are eliminated to achieve maximum inverse association. The symbol r_4^+ denotes the coefficient attained when ties are eliminated to achieve maximum direct association. It should be noticed however that different permutations can lead to r_4^- or r_4^+ and, although more rarely, r_4^- may be equal to r_4^+.

In a summary review of the literature, Amerise & Tarsitano [1], broadly categorizes the methods for resolving ties as single-based and double-based schemes. The former makes use of the information provided by a single ranking, ignoring the possibility of ties in the other ranking. The latter attempts to break ties with a joint procedure involving simultaneously the two rankings. Typical single-based schemes are mid-rank, Dubois and Woodbury methods (see Kendall [12], Amerise & Tarsitano [1]). Double-based techniques are the max-min method and the weighted max-max method.

3.1. Assigning equal ranks

Let $j_{x,i,1}, \cdots, j_{x,i,t_{x,i}}$ be the ranks of the values in the i-th block would have had if there were no ties. According to the Chisini's principle (Chisini [5], Muliere & Parmigiani [15]) the equal ranks method assigns the summary value $\psi_p(j_{x,i,1}, \cdots, j_{x,i,t_{x,i}})$ to each element in the block, where $\psi_p(.)$ is a continuous and strictly increasing function such that the function would remain constant if the summary value were used in place of $j_{x,i,1}, \cdots, j_{x,i,t_{x,i}}$

$$\mu_{x,i,p} = \left[\frac{\sum_{k=1}^{t_{x,i}} (j_{x,i,k})^p}{t_{x,i}} \right]^{1/p}, \quad i = 1, \cdots, m_x. \tag{21}$$

If $p = 1$, expression (21) yields the mid-rank method, which attributes to tied scores the arithmetic average of the ranks which they cover. Regardless of the number of ties, midranks do not change the total mean of the ranks: $(n+1)/2$, but reduce the variance. If $p = 2$, expression (21) yields the quadratic mean. This solution has been proposed by DuBois [7] to preserve the sum of squares of the ranks: $n(n+1)(2n+1)/6$. In this case the condition on the total variance is satisfied, but not that on the total mean, which turns out to be greater than $(n+1)/2$.

Let us consider some special cases suggested by Kendall [12].

- If values are completely tied: $x_{j+1} = x_j$, $y_{j+1} = y_j$, $j = 1, 2, \cdots, n-1$, then rank correlation should indicate perfect direct association. This is true for r_1, but not for r_2, r_3 and r_4, which are zero.

● Suppose that both rankings are the same, that the last member in each is ranked n and that the others are all tied and hence have rank $n/2$ if n is even and rank $(n+1)/2$ if n is odd. Then it will be found that $r_1 = 1$, whereas the other indices are positive (but small), with $r_2 < r_4 < r_3$.

● If one ranking coincides with the identity permutation 1, 2, ⋯, n and has no ties. If the other ranking has the last member ranked n and the others completely tied, we find that $r_1 \to 0.5$, but r_2, r_3, r_4 converge to zero.

These divergences may not be a reason for concern, given that each coefficient compares, in its own way, the degree to which the two variables produce the same ranking of the n observations and react in a different manner in a response to similar situations.

3.2. Max-max method

Gini [9] suggests that although the use of the mid-rank method is well established for many rank correlations, the idea of determining the highest and lowest statistic over the range of possible permutations, within the constraints of tied data, might be beneficial. The same approach is independently followed by Gideon & Hollister [8].

The max-max method uses two special orderings with the aim of determining the pair of permutations that maximizes the direct association or positive correlation and the pair of permutations that maximizes the inverse association or, in absolute terms, the negative correlation. All these diverse tasks have one particular aspect in common: they rely heavily on computational power.

Suppose we regard any set of tied ranks as due to inability on the part of the observer to distinguish real differences; i.e., we assume that there does exist a set of integral ranks although we are ignorant of it on

present evidence. Let $\boldsymbol{Z}_1 = (X, Y)$ be the $n \times 2$ matrix formed by the observed values. Initially, the rows of $\boldsymbol{Z}_1$ are arranged according to increasing magnitude of the first column breaking ties by increasing values of the second column. Untied values in the first column receive the ranks that correspond to their positions. Tied values from j_1 to j_2 are replaced by the permutation of the integers $j_1, \cdots, j_2$, which are generated by increasing values of the second column in the corresponding positions. Unresolved ties (i.e., repeated pairs) are left in their original ordering. Successively, the ties in the positions $j_1, \cdots, j_2$ are substituted by the permutations of $j_1, \cdots, j_2$ which are defined according to decreasing values of the second column. These operations are repeated for all the blocks producing two intermediate permutations of the integers from 1 to n, say (π^*, η^*).

In the second step, the same process is carried out for the matrix $\boldsymbol{Z}_2 = (Y, X, \pi^*, \eta^*)$. In doing so, four final permutations are constructed: (π^+, η^+), which favor direct association most and (π^-, η^-) which favor inverse association most. The max-max estimate of r_h is calculated by using expressions in (2) for the two pairs of extreme permutations

$$r_{4,a} = 0.5\,[r_4(\pi^+, \eta^+) + r_4(\pi^-, \eta^-)]. \tag{22}$$

The subscript a refers to “average”.

According to Gini [9], an improved estimation of r_h in case of ties might be gained by using the weighted max-max method

$$\hat{r}_{4,g} = \theta_1 r_h^- + \theta_2 r_4^+ \quad \text{with} \quad 0 \le \theta_1, \theta_2 \le 1;\ \theta_1 + \theta_2 = 1. \tag{23}$$

If θ_1 is large, then $\hat{r}_{4,g}$ will be close to the maximum inverse association. Similarly, if θ_2 is large, then $\hat{r}_{4,g}$ will be close to the

maximum direct association. Finally, if $\theta_1 = \theta_2 = 0.5$, then $\hat{r}_{4,g} = \hat{r}_{4,a}$. In order to determine the weights, Amerise & Tarsitano [1] suggest that θ_1 and θ_2 can be determined by using a small sample of ν pairs of permutations ($\nu = 1000$ is suggested). Specifically, let $r_{4,u}$ be the rank correlation associated with the u-th pair of randomly chosen permutations for $u = 1, 2, \cdots, \nu$. When the two variables X and Y are highly positively correlated, the ranks of their values tend to be close within most permutations even in those cases where no decision could be made about the ordering. This implies that $r_{4,u} \to r_4^+$. On the other hand, when X and Y are highly negatively correlated, the expected disagreement between the ranks they would have if they were not tied, converges toward the maximum inverse association, that is $r_{4,u} \to r_4^-$. These empirical truths can be exploited to find the relative importance of the bounds in the composition of $r_{4,g}$. To this end, we consider the average distances

$$d_1 = \frac{1}{\nu}\sum_{u=1}^{\nu}|r_{4,u} - r_4^-|, \quad d_2 = \frac{1}{\nu}\sum_{u=1}^{\nu}|r_{4,u} - r_4^+| \tag{24}$$

indicative of a shift from $r_{4,a}$ towards r_4^- or towards r_4^+, respectively. To take these tendencies into account, the weights θ_1 and θ_2 should be computed as follows

$$\theta_1 = \frac{d_1^{-1}}{d_1^{-1} + d_2^{-1}}, \quad \theta_2 = \frac{d_2^{-1}}{d_1^{-1} + d_2^{-1}}. \tag{25}$$

As d_1 decreases (increases), more (less) weight is given to the lower (upper) bound r_4^-, see Niven & Deutsch [16].

3.3. Simulation study

We refer to a scenario in which equal values are due to rounding off of continuous variables. In other words, differences exist, but the computing device or the measuring/recording mechanism fails to distinguish them. We generate $n \in \{12, 18, 24\}$ pairs of values (X, Z) which are uniformly distributed between -6 and 6 with mean zero, variance $\sigma^2(X) = \sigma^2(Z) = 12$ and Pearson product moment correlation, r_0. The values of r_0 belonging to $\{-0.75, -0.50, -0.25, 0, 0.25, 0.50, 0.75\}$.

If we set $Y = r_0 X + (1 - r_0^2)^{0.5} Z$, then it is easily verified that $Cor(X, Y) = r_0$. The generation process produces values $\{\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_n\}$ and $\{\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_n\}$ with an accuracy sufficient to guarantee that no exact ties occur within the precision machine tool limits. Coefficient $\tilde{r}_0$ computed on unrounded values act as the benchmark for approximated rank correlations under ties among the rankings. The values in the pairs $(\tilde{x}_i, \tilde{y}_i)$, $i = 1, 2, \cdots, n$ are then rounded off to the nearest whole number

$$x_i = \text{sgn}(\tilde{x}_i)\lfloor|\tilde{x}_i| + 0.5\rfloor, \quad y_i = \text{sgn}(\tilde{y}_i)\lfloor|\tilde{y}_i| + 0.5\rfloor. \tag{26}$$

For each pairs that passed the filter, we compare three estimates of the coefficients r_h, $h = 1, 3, 4$ obtained in the cases of ties broken according to: mid-rank, max-max and weighed max-max methods. As a criterion for accuracy over $L = 10'000$ replications, we used the absolute mean deviation:

$$AMD_{h,j} = \left[L^{-1} \sum_{l=1}^{L} |\hat{r}_{h,l,j} - r_0| \right], \quad h = 1, 3, 4;\ j = 1, \cdots, 3, \tag{27}$$

which represents the "city-block" distance between observed coefficients and the benchmark r_0. We have excluded Kendall's τ because r_2 it has its own intrinsic system for dealing with ties, which makes the numerator of r_2 invariant when applied to tied rankings, whatever method of breaking ties is employed.

In Table 4, we observe that AMD decreases for both increasing r_0 (in absolute terms) and increasing n; thus, although the techniques for resolving ties have a different conformational orientation, they all converge on r_0. Overall, there are no significant differences between the absolute mean deviations concerning the various coefficients. Nevertheless, with a closer look, we find that the accuracy for max-max methods is marginally-tomoderately lower than that of the mean-rank procedure in the case of r_1 and r_4. In particular, the max-max method appears to be the most effective procedure for estimating r_1 and r_4 in the presence of ties. In contrast, max-max methods are manifestly inadequate in the case of r_3. We interpret this to mean that the extreme values of Gini's cograduation index differ too much to be meaningfully combined by the proposed averages.

Table 4. Absolute mean deviations over $L = 10{,}000$ samples

			True product moment correlation (r_0)						
n	r		− 0.75	− 0.5	− 0.25	0.00	0.25	0.5	0.75
12	r_1	m-r	0.125	0.196	0.231	0.241	0.224	0.186	0.109
		gh	0.116	0.190	0.227	0.239	0.224	0.188	0.115
		wgh	0.115	0.191	0.229	0.240	0.225	0.189	0.114
	r_3	m-r	0.189	0.206	0.202	0.199	0.200	0.204	0.188
		gh	0.467	0.330	0.174	0.099	0.173	0.328	0.466
		wgh	0.469	0.331	0.174	0.098	0.173	0.329	0.468
	r_4	m-r	0.133	0.170	0.207	0.224	0.204	0.169	0.133
		gh	0.111	0.175	0.224	0.243	0.219	0.175	0.110
		wgh	0.109	0.176	0.225	0.245	0.221	0.175	0.108
18	r_1	m-r	0.094	0.157	0.182	0.195	0.181	0.150	0.085
		gh	0.089	0.152	0.179	0.193	0.181	0.153	0.090
		wgh	0.088	0.153	0.180	0.194	0.182	0.153	0.089
	r_3	m-r	0.173	0.180	0.164	0.160	0.166	0.181	0.177
		gh	0.461	0.327	0.169	0.080	0.168	0.327	0.464
		wgh	0.463	0.328	0.169	0.079	0.169	0.328	0.465
	r_4	m-r	0.118	0.134	0.160	0.179	0.162	0.134	0.120
		gh	0.084	0.138	0.176	0.199	0.179	0.137	0.086
		wgh	0.083	0.139	0.178	0.201	0.181	0.138	0.085
24	r_1	m-r	0.081	0.135	0.156	0.167	0.155	0.128	0.071
		gh	0.077	0.131	0.154	0.165	0.155	0.131	0.076
		wgh	0.075	0.131	0.155	0.166	0.156	0.131	0.075
	r_3	m-r	0.173	0.169	0.145	0.137	0.146	0.168	0.171
		gh	0.462	0.326	0.167	0.068	0.167	0.326	0.461
		wgh	0.463	0.327	0.168	0.068	0.168	0.326	0.462
	r_4	m-r	0.115	0.115	0.135	0.153	0.136	0.114	0.114
		gh	0.072	0.118	0.153	0.173	0.154	0.117	0.072
		wgh	0.070	0.119	0.155	0.175	0.156	0.118	0.070

3.4. Application to real data

The graphs reported in Figure 2 show the relationship between the original values and the ranks of log-luminosity of a set of stars and their effective log-temperatures at the surface (Hertzspung-Russell diagram). There is a conspicuous presence of tied values and some records are duplicated: units 2 and 4 and units 33 and 38. We considered redundant the duplicate records and removed them from the data set, prior to analysis. The final number of cases is $n = 45$.

The majority of the stars seems to follow a steep band (i.e., they lie on the main sequence), but there is a small cluster formed by four giant stars (unit 10, 19, 29, and 33) that stand far apart from the rest of the points (see Rousseeuw & Leroy [18])[p. 27-29]. Two other stars (6 and 8) are also far from majority's direction. After a rank transformation of the data, unit 8 remains a suspect outlier, but unit 6 turns to be normal.

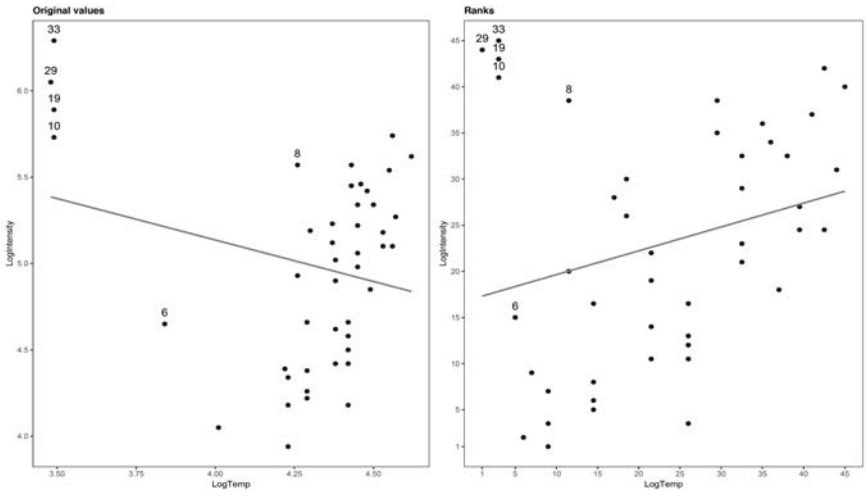

Figure 2. Relationship between luminosity of stars and their effective temperatures.

In Table 5, we summarize rank-correlation coefficients and *p*-values, which are computed using two different treatments of ties: mid-rank (m-r) and Gideon-Hollister (gh). The *p*-values are based on the standard Gaussian distribution. The computation is repeated after excluding the giant stars. The four outliers exert enough influence to distort the correlation between the variables. In fact, with these included in the computation, there is a negative Pearson correlation: $r_0 = -0.247$ and a negative slope of the least square line. The correlation, however, is not impressive as, even at a 5% level, the hypothesis of independence would be rejected. In contrast, when outliers are removed, a highly significant $r_0 = 0.585$ is achieved and the slope of the regression line becomes positive.

The behaviour of the Spearman's r_1 is similar to that of r_0, though the hypothesis of independence is now rejected at the 10% in presence of ties. The same happens to r_2, but with an important difference. Kendall coefficient indicates a positive and relevant (*p*-value less than 5%) relationship between luminosity and temperature either in presence or in absence of the giant stars. This is not surprising because it is well known the inelasticity of Kendall's r_2 at moderate variations in the ranks. See, for example, Tarsitano & Amerise [20]. Notice that r_2 does not change because the group of outliers is not affected by ties.

Table 5. Test of independence

Outliers	Ties	r_0	p-value	r_1	p	r_2	p	r_3	p	r_4	p
Included	m-r	– 0.247	5.1%	0.260	10.1%	0.226	3.7%	0.242	5.8%	0.019	90.4%
	gh			0.257	10.4%	0.226	3.7%	0.119	34.8%	0.012	93.7%
Excluded	m-r	0.585	0.0%	0.658	0.0%	0.478	0.0%	0.529	0.0%	0.678	0.0%
	gh			0.655	0.0%	0.478	0.0%	0.264	3.8%	0.683	0.0%

The values of Gini's r_3 resemble those of Kendall's r_2, but the probability of rejection of independence for the former is slightly higher than for the latter. This is true if ties are dealt with the mid-rank method, but it is not true anymore if ties are broken with the Gideon-Hollister method since, in this case, the observed correlation is halved. Gini's index seems to be strongly reactive to the method used to break ties.

Coefficient r_4 shows the lowest sensitivity to an alteration in the rankings. In absence of outliers, the *p*-value of the statistic changes from 90.4% to 0.002%; in presence of outliers, r_4 gives the indication that no relationships exists between the luminosity and the temperature at surface of the stars. Although it may appear excessively drastic, especially in view of what is shown in the graphs of Figure 2, the proximity of the *p*-value to 100%, in the case of complete data, is a clear and understandable signal that, if a link exists, then it is deeply buried in the noise.

4. Conclusion

The purpose of this paper is to fully explore the sampling behaviour of r_4 when data are affected by ties. The simulations performed allow us to establish that the max-max method, brought to the general attention of researchers and practitioners by Gideon & Hollister [8], yields the estimate of r_4 that best approximates the true Pearson product-moment correlation when equal values are present.

We have identified two favorable features of r_4: the inelasticity respect to changes in the treatment of ties and a high resolution across the $[-1, 1]$ interval that render r_4 a credible measure of rank-order association when a high resolution is needed to differentiate between permutations. These characteristics do not disappear in cases of ranking

with ties. It must be acknowledged that robustness against outliers and against tied values is not without costs. One of the potential costs is the possibility that r_4 can fail to detect weak, but convincing relationships in complex data sets.

Finally, we note that the intuitive idea of a rank correlation computed as a weighted average between the two extremes of the max-max method still deserve to be pursued further.

References

[1] I. L. Amerise and A. Tarsitano, Correction methods for ties in rank correlations, Journal of Applied Statistics 42(12) (2015), 2584-2596.

DOI: https://doi.org/10.1080/02664763.2015.1043870

[2] I. L. Amerise, M. Marozzi and A. Tarsitano, Pvrank: Rank Correlations, R Package Version 1.1.2, 2016.

http://CRAN.R-project.org/package=pvrank

[3] A. Tarsitano and I. L. Amerise, Effectiveness of rank correlations in curvilinear relationships, Behaviormetrika 44(2) (2017), 351-368.

DOI: https://doi.org/10.1007/s41237-017-0020-1

[4] A. D. Barbour and L. H. Y. Chen, The permutation distribution matrix correlation statistics, In: A. D. Barbour and L. H. Y. Chen (Editors), Stein's Method and Applications, Singapore University Press (2005), 223-245.

[5] O. Chisini, Sul Concetto di Media, Periodico di Matematiche 9(2) (1929), 106-116.

[6] D. M. Cifarelli, P. L. Conti and E. Regazzini, On the asymptotic distribution of a general measure of monotone dependence, The Annals of Statistics 24(3) (1996), 1386-1399.

DOI: https://doi.org/10.1214/aos/1032526975

[7] P. H. DuBois, Formulas and tables for rank correlation, The Psychological Record 3 (1939), 46-56.

[8] R. A. Gideon and A. Hollister, A rank correlation coefficient resistant to outliers, Journal of the American Statistical Association 82(398) (1987), 656-666.

[9] C. Gini, Sulla determinazione dell'indice di cograduazione, Metron 13 (1939), 41-48.

[10] G. Girone, S. Montrone and D. Leogrande, La distribuzione campionaria dell'indice di cograduazione di Gini per dimensioni campionarie fino a 24, Annali del Dipartimento di Scienze Statistiche "Carlo Cecchi", Università degli Studi di Bari 24 (2010), 246-271.

[11] W. Hoeffding, A non-parametric test of independence, The Annals of Mathematical Statistics 19(4) (1948), 546-557.

DOI: https://doi.org/10.1214/aoms/1177730150

[12] M. G. Kendall, The treatment of ties in ranking problems, Biometrika 33(3) (1945), 239-251.

DOI: https://doi.org/10.1093/biomet/33.3.239

[13] W. Maciak, Exact null distribution for $n \leq 25$ and probability approximations for Spearman's score in an absence of ties, Journal of Nonparametric Statistics 21(1) (2009), 113-133.

DOI: https://doi.org/10.1080/10485250802401038

[14] A. Mango, A distance function for ranked variables: A proposal for a new rank correlation coefficient, Metodološki Zvezki 3(1) (2006), 9-19.

[15] P. Muliere and G. Parmigiani, Utility and means in the 1930s, Statistical Science 8(4) (1993), 421-432.

DOI: https://doi.org/10.1214/ss/1177010786

[16] E. B. Niven and C. V. Deutsch, Calculating a robust correlation coefficient and quantifying its uncertainty, Computers & Geosciences 40 (2012), 1-9.

DOI: https://doi.org/10.1016/j.cageo.2011.06.021

[17] M. Panneton and P. Robillard, Algorithm AS 54: Kendall's S frequency distribution, Journal of the Royal Statistical Society, Series C 21(3) (1972), 345-348.

DOI: https://doi.org/10.2307/2346291

[18] P. J. Rousseeuw and A. M. Leroy, Robust Regression and Outlier Detection, John Wiley & Sons, New York, 1987.

[19] A. Tarsitano and R. Lombardo, A coefficient of correlation based on ratios of ranks and anti-ranks, Jahrbucher für Nationalökonomie und Statistik 233(2) (2013), 206-224.

DOI: https://doi.org/10.1515/jbnst-2013-0205

[20] A. Tarsitano and I. L. Amerise, On a new measure of rank-order association, Journal of Statistical and Econometric Methods 4(2) (2015), 1-4.

[21] L. Zhao, Z. Bai, C. C. Chao and W.-Q. Liang, Error bound in a central limit theorem of double-indexed permutation statistics, The Annals of Statistics 25(5) (1997), 2210-2227.

DOI: https://doi.org/10.1214/aos/1069362395

[22] Z. Zhang, Quotient correlation: A sample based alternative to Pearson's correlation, The Annals of Statistics 36(2) (2008), 1007-1030.

DOI: https://doi.org/10.1214/009053607000000866

■