International Journal of Statistics and Probability; Vol. 6, No. 2; March 2017
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

A Model to Approximate the Distribution of Rank Order Associations

Agostino Tarsitano[1] & Ilaria L. Amerise[1]

[1] Dipartimento di Economia, Statistica e Finanza - Università della Calabria Via Pietro Bucci, Cubo 1c, 87036 Rende (CS). Italy

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

Received: October 20, 2016 Accepted: December 21, 2016 Online Published: February 21, 2017

doi:10.5539/ijsp.v6n2p66 URL: https://doi.org/10.5539/ijsp.v6n2p66

Abstract

The relationship between two set of ranks can be evaluated by several coefficient of rank-order association. To judge the significance of an observed value of one of these statistics we need a reliable procedure for determining the p-value of the test. In several works the t-Student has been suggested as being relevant for the description of the null distribution of many coefficients. In this article, we propose a new model of density function, the generalized Gaussian on a finite range, which can be used to model data exhibiting a symmetrical unimodal density with a bounded domain. Several simulations illustrate the advantages of this technique over conventional methods. This is particularly useful in the case the number of ranks is larger than the threshold for which the exact null distribution is known, but lower than the threshold for which the asymptotic Gaussian approximation becomes valid.

Keywords: exact p-values, fitting distributions, test of independence

1. Introduction

The extent of agreement between two rankings of n items, numbered from 1 to n, can be tested by using a non-parametric statistic of rank correlation in place of the Pearson product-moment correlation. The most well known statistics of this type are the Spearman, Kendall and Gini coefficients, which will be denoted, respectively, as r_1, r_2 and r_3. The former is the most often used measure in research in which the dependence is assumed monotonic but otherwise arbitrary. In comparison with r_1 and r_2, Gini's r_3 seems to be applied rather rarely at present, although its characteristics are similar, and sometimes better, to those of the other rank correlations.

To judge the significance of an observed value of one of these statistics, say r_h, we need the exact distribution of r_h under the hypothesis of independence or, at least, a reliable procedure for determining the p-value of the test. Significance levels could, for example, be calculated using asymptotic methods. In this regard, the convergence to the Gaussian distribution renders its use legitimate in interpolating the p-values of r_h, $h = 1, 2, 3$. Nonetheless, already Old (Olds, E. G., 1938) states that a distribution with a finite range causes trouble at the tails when a Gaussian fit is attempted, and, this is particularly relevant to studies where we are particularly interested in the tails. KIendall *et al.* (KIendall, *et al.*, 1939) add that Gaussian approximation is satisfactory for moderately large values, but for small values it is subject to the disadvantage inherent in any attempt to represent a distribution of finite range by one of infinite range, that is, the fit near the tails it is not likely to be very good. On the other hand, rank correlation statistics lies in the interval from -1 to 1 and we think it is better for clarity to test them by using a theoretical curve with a bounded, rather than infinite, domain.

In order to circumvent these difficulties, many researchers have looked for probability densities which are capable of fitting the distribution of rank correlations appropriately, including: Johnson S_B (Johnson, N-L., 1949), Tadikamalla-Johnson L_B (Tadikamalla & Johnson, N-L., 1994). In particular, Pitman (Pitman, E. J. G., 1937) noted that the first four moments of the r_1-distribution were similar to the first four moments of the symmetrical beta or Pearson type II distribution. Continuing this idea, Landenna *et al.* (Landenna, G., 1989) proposed the symmetrical beta for the Gini coefficient r_3 and Vittadini (Vittadini, G., 1996) suggested it for the Kendall coefficient r_2. One key factor behind the wide diffusion of this model is the strict relationship between the symmetrical beta curve and the Student's t density function (Willink, 2009). This allows for the use of easy tables and hence ensures computational convenience and simple checking of results.

Our objective in this paper is to devise a new model for estimating the p-values of some rank association indices in the case n is larger than the threshold for which the exact null distribution is known but lower than the value of n for which the Gaussian approximation becomes valid. The structure of the paper is as follows. In Section 2, we succinctly discuss the characteristics of r_1, r_2 and r_3. A new density function, the generalized Gaussian on a finite range (GGFR), is introduced in Section 3 and the prediction of p values is presented in Section 4. We conclude in Section 5.

2. Indices of Rank Order Association

The degree of monotone association between rankings can be measured by rank index of association. The coefficients reported in Table 1 are in general use.

Table 1. Rank association statistics

Coefficient	Formula	Number of distinct values
Spearman	$r_1 = \left(\frac{3}{n^3 - n}\right) S_1$	$\left(n^3 - n\right)/6 + 1$
Kendall	$r_2 = \left(\frac{2}{n^2 - n}\right) S_2$	$\left(n^2 - n\right)/2 + 1$
Gini	$r_3 = \left(\frac{4}{n^2 - k_n}\right) S_3,\ k_n = n \mod 2$	$\left(n^2 - k_n\right)/2 + 1,$

Here $S_1 = \sum_{i=1}^{n}\left[(n+1-i-\pi_i)^2-(i-\pi_i)^2\right]$ where π is a permutation of order n. $S_2 = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} sign\left(\pi_j-\pi_i\right)$ where sign(x) takes the values $-1, 0, +1$ according to whether x is negative, zero or positive, $S_3 = \sum_{i=1}^{n} (|n+1-i-\pi_i|-|i-\pi_i|)$. We remark that the expressions presented in Table 1 assume absence of ties.

Coefficients in Table 1 vary within the range:$[-1, 1]$. The extremes are achieved if and only if there is perfect association, negative or positive, for all pairs:$r_h(\eta, \eta) = r_h(\pi, \pi) = 1$, $r_h(\eta, \eta^*) = r_h(\pi, \pi^*) = -1$ where $\pi^* = n+1-\pi$ and $\eta^* = n+1-\eta$ are the reverse permutations of π and η. The larger r_h is, ignoring the sign, the stronger the association between rankings is. All the three indices can be interpreted as differences between the distance from perfect direct association $(1, 2, \cdots, n)$ and the distance from perfect inverse association $(n, n-1, \cdots, 1)$.

Van de Wiel & Di Bucchianico (Van de Wiel, M. A., & Di Bucchianico, A., 2001), compute the null distribution of r_1 for $n = 19, \ldots, 22$ using the representation of its probability generating function as a permanent (a signless determinant) with monomial entries. See also Maciak (Maciak, W., 2009). It is interesting to note that he quantities S_1 and S_3 appearing in r_1 and r_3 can be expressed as a sum of parts which allows the use of combinations of sub permutations that significantly reduce the amount of computation required to build the exact distribution. See Otten (Otten, A., 1973) for a division of the permutations in two groups. Girone *et al.* (Girone, G., et al., 2010) went further by breaking up the permutations into four groups and executing a parallel processing scheme that, by the way, is naturally fit to Otten's proposal. Research to date has obtained the null distribution of the Spearman coefficient up to $n = 26$ (Gustavson, 2009) and that of the Gini coefficient for up to $n = 24$. The same procedures cannot be applied for Kendall's S_2, which, however, benefits from a recurrence relationship. See Panneton & Robillard (Panneton & Robillard, 1972).

Under the null hypothesis of independent rankings, the distributions of r_1, r_2 and r_3 are symmetrical and have support in $[-1, 1]$. All the odd moments are zero because of the symmetry. Furthermore, and this is essential in our paper, their variance and kurtosis are known as polynomials in n, as it is shown in Table 2.

Table 2. Second and fourth moments of r_1 r_2 r_3

	$\mu_2(n)$	$\mu_4(n)$
r_1	$\frac{1}{n-1}$	$\frac{3\left(25n^3 - 38n^2 - 35n + 72\right)}{25n(n+1)(n-1)^3}$
r_2	$\frac{2(2n+5)}{9n(n-1)}$	$\frac{100n^4+328n^3-127n^2-997n-372}{1350\,[0.5n(n-1)]^3}$
r_3	$\frac{2}{3(n-1)}\left[\frac{n^2+2+k_n}{n^2-k_n}\right]$	$\frac{4[35n^7-(111-35k_n)n^6+(153+29k_n)n^5-(366-59k_n)n^4+(304+11k_n)n^3-(456-114k_n)n^2-(912-492k_n)n+(1248-933k_n)]}{n^{k_n}(105-2k_n)(n+k_n)^3(n-k_n)^4(n-3+k_n)}$

where $k_n = n \bmod 2$. Both $\mu_2(n)$ and $\mu_4(n)$ are decreasing function of n with the values relative to r_3 always intermediate between those of r_1 and r_2 (the former systematically greater than the latter). It can also be observed that, because of the presence of k_n, the moments of r_3 have an oscillating character due to the odd-even parity of n, that is the number of items.

3. A New Model of Density Function

A good model should reproduce the characteristics of r_1, r_2 and r_3 generally observed over the whole population of permutations. Specifically, curves must be unimodal, symmetrical around zero, bounded in the interval $[-1, 1]$; moreover, as the range widens, they must tend towards the Gaussian probability distribution. The usual procedure to determine theoretical probabilities and expected frequencies is to find a curve capable of providing all the required peculiarities cited above, and then to integrate for the probabilities over the given intervals.

The main contribution of the present paper is to provide a theoretical explanation for the behavior of several coefficients of rank order association. Suppose that the relative variation of the probability density of the absolute value of the random variables $f(|r|)$ representing the rank order association is inversely proportional to $1 - f(|r|)$.

$$\frac{d\left[f(|r|)\right]}{f(|r|)}\frac{1}{d\left[|r|^{\lambda_1}\right]} = -\frac{\lambda_2}{1-|r|^{\lambda_1}} \quad \text{where } \lambda_1 \geq 1. \tag{1}$$

Vianelli (Vianelli, 1968) shows that, under normalization condition, the integration of (1) leads to

$$f(r, \boldsymbol{\lambda}) = \frac{\lambda_1\left[1-|r|^{\lambda_1}\right]^{\lambda_2}}{2B\left(\lambda_1{}^{-1}, \lambda_2+1\right)} \quad \text{with } |r| \leq 1; \quad \boldsymbol{\lambda} = (\lambda_1, \lambda_2), \quad \lambda_1, \lambda_2 > 0. \tag{2}$$

We call this curve GGFR (generalized Gaussian distribution with finite range). It is easily verified that, for $\lambda_1, \lambda_2 > 1$, $f(r, \boldsymbol{\lambda})$ takes a bell shaped form. If $0 < \lambda_1 \leq 1$ and $\lambda_2 > 1$ the density (2) resembles the characteristic shape of the Laplace density (with its sharp peak at the mode). The parameter λ_2 mainly influences the tails of the distribution. The symmetrical beta (alias t-Student) is a special case of (2) for $\lambda_1 = 2$ and $\lambda_2 = 0.5n - 2$. The GGFR density is symmetrical, unimodal with mode at zero, is supported within interval $[-1, 1]$, has two inflection points located at $\pm\left[(\lambda_1 - 1)/(\lambda_1\lambda_2 - 1)\right]^{1/\lambda_1}$. See Vianelli (1983). Furthermore, $f(r, \boldsymbol{\lambda})$ converges towards the Gaussian distribution as the range is widened. See Devroye (Devroye, 1986)[p.433-437].

To estimate the parameters of the GGFR we will follow the moment-matching method as. In this regard, the second and fourth centered moments of the GGFR density are

$$\mu_2(\boldsymbol{\lambda}) = \frac{B\left(3\lambda_1^{-1}, \lambda_2+1\right)}{B\left(\lambda_1^{-1}, \lambda_2+1\right)}; \qquad \mu_4(\boldsymbol{\lambda}) = \frac{B\left(5\lambda_1^{-1}, \lambda_2+1\right)}{B\left(\lambda_1^{-1}, \lambda_2+1\right)}. \tag{3}$$

The variance $\mu_2(\boldsymbol{\lambda})$ increases for a higher λ_1 or for a lower λ_2, whereas the excess kurtosis $\gamma_2(\boldsymbol{\lambda}) = \mu_4(\boldsymbol{\lambda})/\sigma^4(\boldsymbol{\lambda}) - 3$ rises to zero for a decreasing λ_2 or diminishes to zero for an increasing λ_1.

Let us consider a loss function in which the lowest two even moments of $r_{h,n}$ (for n ranks) are matched to those of a GGFR density.

$$G_{n,h}(\boldsymbol{\lambda}) = \text{minimize}_{\boldsymbol{\lambda}}\left\{\max\left\{|\mu_2(\boldsymbol{\lambda}) - \mu_{2,h}(n)|, |\mu_4(\boldsymbol{\lambda}) - \mu_{4,h}(n)|\right\}\right\} \tag{4}$$

where h=1, 2, 3 and $\lambda = (\lambda_1, \lambda_2)$. The GGFR density has two exponential parameters that make (4) highly nonlinear. In addition, the presence of a beta function depending on the unknown parameters can create difficulties in numerical stability. To increase the chances of getting a global solution in reasonable computational times, we executed a controlled random search (CRS) algorithm discussed, for example, in Conlon (Conlon, 1992) and Brachetti *et al.* (Brachetti, et al., 1997). See Amerise *et al.* (Amerise, 2015) for more details on the procedure used.

In Table 3 we show the estimates of $\boldsymbol{\lambda}$ for a few values of n with $G(\boldsymbol{\lambda}) < 0.1 \times 10^{-16}$ in each experiment. The rows σ^2 and γ_2 indicate, respectively, the variance and the excess kurtosis of the coefficients obtained on the basis of Table 2. All the three rank correlations have a platykurtic null distribution, which is flatter than Gaussian. This characteristic is more evident for Spearman's r_1 whereas, under this point of view, Kendall's r_2 is the closest to the Gaussian distribution.

Generally, as n increases, the parameter estimates increase; moreover, the variance of the best fitting density decreases and the associated excess kurtosis remains negative but tends to zero (which could be a symptom of asymptotic Gaussianity). The trend for Gini's cograduation has two branches, one for even and the other for odd parity of n. This alternating behavior is due to the strong effect of $k_n = n \bmod 2$, which appears both in the expressions and in the moments of r_3.

Tbale 3. Estimation of the parameters of the GGFR

		number of ranks							
		21	22	23	24	25	26	27	28
r_1	λ_1	1.8969	1.9031	1.9086	1.9135	1.9179	1.9220	1.9256	1.9290
	λ_2	7.8254	8.3116	8.7984	9.2858	9.7737	10.2621	10.7509	11.2402
	σ^2	0.0500	0.0476	0.0455	0.0435	0.0417	0.0400	0.0385	0.0370
	γ_2	-0.2194	-0.2094	-0.2002	-0.1918	-0.1841	-0.1769	-0.1703	-0.1642
r_2	λ_1	1.9423	1.9466	1.9504	1.9538	1.9568	1.9594	1.9618	1.9639
	λ_2	17.4755	18.5811	19.6883	20.7967	21.9063	23.0169	24.1284	25.2407
	σ^2	0.0249	0.0236	0.0224	0.0213	0.0204	0.0195	0.0187	0.0179
	γ_2	-0.1017	-0.0971	-0.0928	-0.0890	-0.0854	-0.0822	-0.0791	-0.0763
r_3	λ_1	1.9634	2.1549	1.9680	2.1359	1.9716	2.1221	1.9745	2.1117
	λ_2	12.9026	16.5302	14.4071	18.0374	15.9097	19.5723	17.4109	21.1280
	σ^2	0.0336	0.0319	0.0305	0.0291	0.0280	0.0267	0.0258	0.0248
	γ_2	-0.1664	-0.2703	-0.1521	-0.2477	-0.1400	-0.2294	-0.1297	-0.2141

4. Approximations of p-values

To compare the GGFR solution with the asymptotic method (Gaussian density) and t-Student alternative for small samples, we consider both exact and fitted significance levels α of the test H_0 : rankings are independent against H_1 : rankings are dependent, by using ρ_1 (Spearman), ρ_2 (Kendall), ρ_3 (Gini). Let r_1, r_2 and r_3 indicate, respectively, the empirical values of Spearman, Kendall and Gini rank order associations. The statistics involved in the t-Student approximation are

$$r_1^*=r_1\sqrt{\frac{m_1}{1-r_1^2}}\sim t_{m_1},\quad r_2^*=r_2\sqrt{\frac{m_2}{1-r_2^2}}\sim t_{\lfloor m_2\rfloor}\quad r_3^*=r_3\sqrt{\frac{m_3}{1-r_3^2}}\sim t_{\lfloor m_3+0.5\rfloor} \tag{5}$$

$$\text{with}\quad m_1=n-2,\ m_2=\frac{9n\,(n-1)}{(4n+10)}-1,\ m_3=\frac{\left[3\,(n-1)\left(n^2-k_n\right)\right]}{2\,(n^2+2+k_n)}-2.$$

The statistics involved in the standard Gaussian approximation are

$$r_1^+=r_1\sqrt{n-1},\qquad r_2^+=r_2\sqrt{\frac{9n(n-1)}{4n+10}},\qquad r_3^+=r_3\sqrt{1.5n}\,. \tag{6}$$

4.1 Accuracy of Approximations

Iman & Conover (Iman & Conover, 1978) correctly observe that 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 $C_{\alpha,h}$) or the next larger exact size called liberal p-value (denoted by $L_{\alpha,h}$). Let $\rho_{\alpha,h,C}$ and $\rho_{\alpha,h,L}$ be the quantiles of $\rho_h, h=1,2,3$ corresponding to the probability levels $C_{\alpha,h}$ and $L_{h,\alpha}$, respectively. The test of H_0 and H_1 above is conclusive if both the conservative and the liberal p-values lie on the same side with respect of the prefixed nominal level α. If $C_{\alpha,h}<\alpha<L_{\alpha,h}$, then the test is unreliable.

To investigate the accuracy of the proposed approximations, we examine a set of 500 nominal levels $\alpha=0.0001, 0.0002, \cdots, 0.0500$. For each α we compute both the actual $C_{\alpha,h}$ and the fitted $\widehat{C}_{\alpha,h,k}$ conservative p-values and repeat the same calculation for the actual $L_{\alpha,h}$ and the fitted $\widehat{L}_{\alpha,h,k}$ liberal p-values. The fitted p-values are based on GGFR ($k=1$), Gaussian ($k=2$) and t-Student ($k=3$) probability densities. A summary of the results is presented in Table 4. The most notable figures are emphasized in bold font. For reason of space, attention is focused on $n=19,\cdots,24$ which are the largest values of n for which the exact null distribution is known for all the three rank correlations. The quantity

$$\delta_{\alpha,h,k}=0.5\left(\left|\widehat{C}_{\alpha,h,k}-C_{\alpha,h}\right|+\left|\widehat{L}_{\alpha,h,k}-L_{\alpha,h}\right|\right) \tag{7}$$

$k=1,2,3;\ h=1,2,3;\ \alpha\in\{0.0001,\cdots,0.05\}$ gives the average distance between lower and upper fitted and actual significance levels and it is used to assess the quality of approximation. High values of $\delta_{\alpha,h,k}$ indicate that approximations to the null distribution of the rank correlation h, based on model k, far exceed or under-run at least one exact threshold at the level α. Low values of $\delta_{\alpha,h,k}$ point to good approximations. The value $\delta_{\alpha,h,k} = 0$ tends to zero when the gap between exact conservative and liberal p-values tends to vanish and, at the same time, the k-th theoretical model yields an almost exact match to these p-values.

Table 4. Summary statistics for $\delta_{h,\alpha,k} \times 10^3$

		Spearman r_1			Kendall r_2			Gini r_3		
n		GGFR	t	Gauss	GGFR	t	Gauss	GGFR	t	Gauss
19.000	Max	**0.136**	0.838	1.733	**3.291**	3.461	3.720	**2.803**	3.059	8.167
	Median	**0.025**	0.722	0.873	1.956	2.341	**1.662**	**1.673**	2.025	4.266
	St.Dev.	**0.048**	0.147	0.555	0.902	**0.898**	1.179	**0.790**	0.822	2.581
20.000	Max	**0.119**	0.768	1.633	**3.180**	3.282	3.648	**2.665**	2.949	7.498
	Median	**0.035**	0.657	0.823	1.739	2.080	**1.420**	**1.499**	1.870	3.576
	St.Dev.	**0.035**	0.137	0.525	0.833	**0.823**	1.087	**0.736**	0.778	2.325
21.000	Max	**0.106**	0.709	1.544	**2.921**	3.002	3.342	**2.533**	2.531	7.643
	Median	**0.054**	0.599	0.789	1.667	1.960	**1.388**	**1.492**	1.622	3.979
	St.Dev.	**0.027**	0.130	0.498	0.780	**0.764**	1.020	0.683	**0.663**	2.293
22.000	Max	**0.095**	0.658	1.465	**2.578**	2.666	2.891	**2.25**	2.328	6.504
	Median	**0.065**	0.548	0.752	1.519	1.780	**1.232**	**1.365**	1.539	3.441
	St.Dev.	**0.025**	0.124	0.473	0.728	**0.709**	0.953	0.636	**0.631**	2.071
23.000	Max	**0.089**	0.613	1.393	**2.417**	2.481	2.715	**2.167**	2.329	6.698
	Median	**0.070**	0.507	0.720	1.475	1.699	**1.229**	**1.256**	1.510	3.406
	St.Dev.	**0.027**	0.119	0.451	0.683	**0.660**	0.894	**0.596**	0.612	2.058
24.000	Max	**0.099**	0.574	1.328	**2.382**	2.402	2.730	**2.101**	2.287	6.306
	Median	**0.072**	0.470	0.700	1.369	1.567	**1.123**	**1.172**	1.448	3.026
	St.Dev.	**0.030**	0.115	0.431	0.644	**0.618**	0.846	**0.559**	0.584	1.874

The findings in Table 4 demonstrate that GGFR provides a more accurate fitting of the null distribution of r_1 and r_3 than Gaussian or t-Student densities for all the values of n considered in the table. This is not surprising, since it means that the addition of a shape parameter to the symmetrical beta yields a significant enhancement over the conventional method. The findings also show that a density with a bounded domain yields better fittings than those obtained from t-Student and Gaussian distributions, which takes value on the entire real line.

In the case of r_2, GGFR has the lowest maximum error but does not always give the smallest median error. On the contrary, in this respect, the Gaussian model offers a better fit for $n > 19$. The good performance of the Gaussian density might be explained by the low thickness of the tails of the null distribution of Kendall's r_2. In fact, fittings with a Student's t density do not improve the fittings with of the Gaussian density, even though the former yield the smallest standard errors of the criterion (7). We attribute this mixed behavior of GGFR mainly to the varying size of serrations in the frequency polygon of r_2, which are more intense than those in r_1 and r_3.

In general, all the indicators: max, median and standard deviation show a common tendency to decrease as the number of ranks increases, presumably as an effect of the concomitant convergence of the null distribution of r_1, r_2, r_3 to the Gaussian density. To this there are some exceptions, notably in the column headed GGFR. As it can easily be noted, however, the values of the indicators are very small in comparison with those of the other columns so that it may be assumed that the observed deviations from the common decreasing pattern reflect small-scale fluctuations in the computations.

4.2 Practical Applications

To illustrate how the proposed procedure can be applied in practice, Table 5 shows a few examples selected with a view towards demonstrating that the level at which rank correlations were considered statistically significant in those cases was uncertain and/or the conclusions questionable. Note that, in all the cases reported, the authors of the various papers have made use of two-sided tests. Naturally, we have given most attention to those studies in which the number of ranks n are larger than those included in available tables of the exact null distribution of r_1, r_2 and r_3, because we assume that researchers have little motivation to use approximate p-values whenever exact p-values are available (except that in cases in which tests are inconclusive). Examination of Table 5, reveals that the GGFR offers approximations to the p-value of r_1, which are intermediate between those of the t-Student and the Gaussian. For the coefficient r_2, the GGFR is more liberal than the other two. Overall, it appears that a single measure often does not appropriately reflect the strength of the

Table 5. Observed and predicted p-values.

		Observed		Nominal	Approximate Sig. Lev.		
Reference	Coefficient	statistic	n	Sig. Lev.	GGFR	t-Student	Gaussian
Chong *et al.* (2004)	r_1	-0.423	59	0.0010	0.0009	0.0008	0.0013
Jones *et al.* (2012)	r_1	-0.240	121	0.0100	0.0081	0.0080	0.0086
Thornton (1943)	r_1	-0.080	44	0.6200	0.6034	0.6057	0.5999
Dupuis *et al.* (1995)	r_2	0.240	90	0.0010	0.0014	0.0007	0.0008
Swann *et al.* (2008)	r_2	0.169	74	0.050	0.0364	0.0327	0.0332
Merry *et al.* (2008)	r_2	0.245	76	0.0024	0.0017	0.0016	0.0017
Coccia & Rolfo (2008)	r_3	-0.492	33	0.0010	0.0002	0.0003	0.0005
Salvemini (1951)	r_3	0.340	30	0.0232	0.0119	0.0233	0.0226
Amato (1951)	r_3	0.204	91	0.0178	0.0123	0.0172	0.0172

association so that the combined evaluation of more than one approximation may throw light on the correct significance of a test. With regard of r_3, the p-values proposed by GGFR are more conservative than Gaussian and t-Student distributions.

5. Discussion and Conclusion

Over recent years, the number of ranks for which the exact null distribution is fully available has increased for many measures of monotone association. For problems involving a number of ranks, which is not included in the existing software though, it is necessary to resort to the omnipresent Gaussian approximations while awaiting faster and more economical computers. However, the Gaussian density can be misleading, particularly in the tails, which often are the most important part. In this paper, we have demonstrated the usefulness of the generalized Gaussian density with finite range (GGFR) for fitting the exact null distribution of three statistics which are routinely used for measuring the correlation between two rankings: Spearman, Kendall and Gini coefficients. All that is required is that variance and kurtosis be known functions of the number of ranks.

The performance of the GGFR is decidedly superior to that of t-Student and Gaussian distributions, which are traditionally employed to estimate tail probabilities for the Spearman and the Gini coefficients. The situation regarding the Kendall coefficient is rather different. In this case, the Gaussian model achieves the best results. Improvement over conventional procedures (Gaussian and t-Student densities) does not appear impressive, but touches on the distribution tails, which are the most interesting from a practical point of view. It must be added that the GGFR achieves the largest improvement in fitting the null distribution of Spearman's ρ, which is the most known and probably most used rank correlation coefficient.

References

Amato, V. (1951). Sulla distribuzione dell'indice di cograduazione di Gini. *Statistica*, (14), 515-519.

Amerise, I. L., Marozzi, M., & Tarsitano, A. (2015). Pvrank: rank correlations. *R package version 1.1.* http://CRAN.R-project.org/package=pvrank.

Brachetti, P., Ciccoli, M., Di Pillo, G., & Lucidi, S. (1997). A new version of the Price's algorithm for global optimization. *J. Global. Optim.*, (10), 165-184. https://doi.org/10.1023/A:1008250020656

Chong, A. Y., et al. (2004). Endothelial dysfunction and damage in congestive heart failure: relation of flow-mediated dilation to circulating endothelial cells, plasma indexes of endothelial damage, and brain natriuretic peptide. *Circulation*, (110), 1794-1798. https://doi.org/10.1016/j.amjcard.2005.09.113

Coccia, M., & Rolfo, S. (2008). Strategic change of public research units in their scientific activity, *Technovation*, (28), 485-494. https://doi.org/10.1016/j.technovation.2008.02.005

Conlon, M. (1992). The controlled random search procedure for function optimization. *Comm. in Stat. - Sim. and Comp.*, (21), 912-923. https://doi.org/10.1080/03610919208813057

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. New York, NY: Springer-Verlag. https://doi.org/10.1007/978-1-4613-8643-8

Dupuis, L. A., et al. (1995). Relation of terrestrial-breeding amphibian abundance to tree-stand age. *Conserv. Biol.*, (9), 649-653. https://doi.org/10.1046/j.1523-1739.1995.09030645.X

Girone, G., Montrone, S., & Leogrande, D. (2010). 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), 246-271.

Gustafson, L. (2000). Spearman rho null distribution. Available at http://www.luke-g.com/math/spearman/index.html

Iman, L., & Conover, W. J. (1978). Approximations of the critical region for Spearman's rho with and without ties present. *Comm. in Stat. - Sim. and Comp.*, (7), 269-282. https://doi.org/10.1080/03610917808812076

Johnson, N-L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, (36), 149-176. https://doi.org/10.2307/2332539

Jones, P. W., et al. (2012). Tests of the responsiveness of the COPD assessment test following acute exacerbation and pulmonary rehabilitation. *Chest*, (142), 134-140. https://doi.org/10.1378/chest.11-0309

Kendall, M. G., Kendall, S. F. H., & Babington Smith, B. (1939). The distribution of Spearman's coefficient of rank correlation in a universe in which all rankings occur an equal number of times. *Biometrika*, (30), 251-273. https://doi.org/10.1093/biomet/30.3-4.251

Kishimoto, et al. (2010). The relationship between the size of caudolateral curvilinear osteophyte of the canine femoral neck and the radiographic view. *J. Vet. Sci.*, (11), 89-91. https://doi.org/10.41.42/jvs.2010.11.1.89

Landenna, G., Scagni, A. & Boldrini, M. (1989). An approximated distribution of the Gini's rank association coefficient. *Comm. in Stat. Theor. and Meth.*, (18), 2017-2026. https://doi.org/10.1080/03910928908830019

Maciak, W. (2009). Exact null distribution for $n \leq 25$ and probability approximations for Spearman's score in an absence of ties. *J. Nonpar. Statist.*, (21), 113-133.

Merry, B., Gallotta, M. & Hultquist, C. (2008). Challenges in running a computer olympiad in South Africa. *Olympiads in Informatics*, (2), 115-114.

Olds, E. G. (1938). Distributions of sums of squares of rank differences for small numbers of individuals. *Ann. Math. Statist.*, (9),133-148. https://doi.org/10.1214/aoms/1177732332

Otten, A. (1973). The null distribution of Spearman's ρ when $n = 13\,(1)\,16$. *Stat. Neer.*, (27), 19-20. https://doi.org/10.1111/j.1467-9574.1973.tb00204.x

Panneton, M., & Robillard, P. (1972) . Algorithm AS54: Kendall's S frequency distribution. *App. Stat.*, (21), 345-348. https://doi.org/10.2307/2346291

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations. II. the correlation coefficient test. *J. Royal Stat. Soc.*, (4), ?25-233. https://doi.org/10.2307/2983647

Team, R. C. (2013). *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Salvemini, T. (1951). Sui vari indici di cograduazione. *Statistica*, (2), 133-154.

Swann, A. C., et al. (2008). Impulsivity: Differential relationship to depression and mania in bipolar disorder, *J. of Aff. Dis.*, (106), 241-248. https://doi.org/10.1016/j.jad.2007.07.011

Tadikamalla, P. R., & Johnson, N. L. (1994). Tables to facilitate fitting Tadikamalla and Johnson's LB distributions II. Both terminals fixed. *Comm. in Stat. - Sim. and Comp.*, (23), 569-581. https://doi.org/10.1080/03610919408813187

Thornton, G. R. (1943). The significance of rank difference coefficients of correlation. *Psychometrika*, (8), 211-222. https://doi.org/10.1007/BF02288705

Van de Wiel, M. A., & Di Bucchianico, A. (2001). Fast computation of the exact null distribution of Spearman's ρ and Page's L statistic for samples with and without ties. *J. Stat. Plan. Inference*, (92), 133-145. https://doi.org/10.1016/S0378-3758(00)00166-X

Vianelli, S. (1968). Sulle curve normali di ordine r per intervalli finiti delle variabili statistiche. *Annali della Facoltà di Economia e Commercio dellUniversità di Palermo*, (2).

Vianelli, S. (1983). The family of normal and lognormal distributions of order r. *Metron*, (41), 3-10.

Vittadini, G. (1996). Una famiglia di distribuzione per i test di associazione. *In Atti della XXXVIII riunione scientifica della S.I.S.*, Rimini, Italy 9-13 Aprile, (2), 521-528.

Willink, R. (2009). A single form for t-distributions and symmetric beta distributions. *Comm. in Stat. Theor. and Meth.*, (39), 170-176. https://doi.org/10.1080/03610920802650346

Copyrights