6-1987

A Rank Correlation Coefficient Resistant to Outliers

Rudy Gideon
University of Montana, Missoula

Robert Ashley Hollister
The University of Montana

A Rank Correlation Coefficient Resistant to Outliers

RUDY A. GIDEON and ROBERT A. HOLLISTER*

In this article, a nonparametric correlation coefficient is defined that is based on the principle of maximum deviations. This new correlation coefficient, R_g, is easy to compute by hand for small to medium sample sizes. In comparing it with existing correlation coefficients, it was found to be superior in a sampling situation that we call "biased outliers," and hence appears to be more resistant to outliers than the Pearson, Spearman, and Kendall correlation coefficients. In a correlational study not included in this article of some social data consisting of five variables for each of 51 observations, R_g was compared with the other three correlation coefficients. There was agreement on 8 of the 10 possible correlations, but in one case, R_g was significant when the others were not, and in yet another case, R_g was not significant when the others were. A further analysis of this data set indicated that there were three to six data points that were anomalies and had a severe effect on the other correlations but not R_g. Apparently, the statistic R_g measures association in a unique fashion. This different measure of association for real data is extended to a population interpretation and expressed in terms of the copula function.

In consideration of ties, this article suggests a randomization method and a computation of the minimum and maximum possible correlation values when ties are present. These ideas are illustrated with an example.

Critical values of R_g and enough examples are included so that this new statistic can be applied to data. The success that we have had with the use of R_g in hypothesis testing suggests that R_g may have important applications wherever robustness is desired.

KEY WORDS: Permutation group; Copula function; Simulated distribution; Robust rank correlation coefficients; Independence testing; Outliers and their effect on correlation coefficients.

1. INTRODUCTION

Some sampling situations involve bivariate data that look correlated but have one or more data points that appear inconsistent with the bulk of the data. The trimmed mean has been suggested as an appropriate procedure in certain estimation problems. In some data, however, the "outlier" part of the data is in fact reliable data and should not be excluded. The proposed correlation coefficient is not as sensitive to inconsistent data as the most commonly used ones.

The data shown in Figure 1 were observed in a YMCA fourth and fifth grade boys' basketball league in Missoula, Montana in 1979. The won–lost standings for the 16-team league are given as well as a sportsmanship ranking that was an accumulation of a subjective evaluation after each game.

In general, we see that the better teams had poorer sportmanship rankings, except for the fourth and thirteenth best teams. In evaluating this relationship one would desire a correlation coefficient that illuminates the general relations and is not unduly influenced by several possibly unusual but yet accurate data. Let us compute the Spearman R_s (1904), the Kendall R_k (1938), the quadrant R_q (Blomqvist 1950), and the new correlation coefficient, denoted by R_g, for the data in Figure 1 and for two perturbations of this data: (a) when the sportsmanship rankings of teams 4 and 13 are interchanged (more consistent); and (b) when teams 4 and 13 are left as they were observed, but the sportsmanship rankings of the best and worst (first and sixteenth) teams are interchanged (less consistent). The results are given in Table 1.

It can be seen that the greatest changes in the values of the correlation coefficients over the three cases occur in the existing correlations and that R_g changes least. This is backed up by computation of the corresponding one-sided probability values for each result. This resistance-to-change property of R_g and the corresponding probability values are possibly of great value in detecting relationships between variables that are masked by current correlation coefficients.

2. DEFINITION OF CORRELATION COEFFICIENT R_g

Let $\mathbf{p} = (p_1, p_2, \ldots, p_N)$ be a permutation of the first N positive integers. For a bivariate set of data $(x_i, y_i)_{i=1}^N$, let $r(x_i)$ be the rank of x_i among the x data and similarly define $r(y_i)$. We shall assume a continuous distribution so that with probability 1 the ranks are unique. Now order the x data and let p_i be the rank of the y datum that corresponds to the ith smallest x value. In the YMCA example in Figure 1 with the won–lost ranks as the x values and the sportsmanship ranks as the y values, this vector $\mathbf{p} = (14, 11, 16, \ldots, 5)$ appears above the x-axis.

Let S_N be the symmetric group of degree N. There are $N!$ possible $\mathbf{p}$ in S_N. Let the group operation $\circ$ be the composition of mappings. Thus if both $\mathbf{p} = (p_1, p_2, \ldots, p_N)$ and $\mathbf{q} = (q_1, q_2, \ldots, q_N)$ are in S_N, then $\mathbf{p} \circ \mathbf{q}$ has for its ith component $\mathbf{p} \circ \mathbf{q}(i) = p_{(q_i)}$ $(i = 1, 2, \ldots, N)$. For each (X, Y) data set of size N, permutation p is denoted by $\mathbf{p} = \mathbf{p}(X, Y)$ and formally defined by $p_{r(x_i)} = p(r(x_i)) = r(y_i)$, where (x_i, y_i) is the ith pair in the data set $(i = 1, 2, \ldots, N)$.

There are two permutations in S_N that are of special interest. They are the *identity permutation*, $\mathbf{e} = (1, 2, \ldots, N)$, and the *reverse permutation*, $\boldsymbol{\epsilon} = (N, N - 1, \ldots, 1)$. Since $\boldsymbol{\epsilon}(i) = N + 1 - i$, $\boldsymbol{\epsilon} \circ \mathbf{p} = (N + 1 - p_1, \ldots, N + 1 - p_N)$ and $\mathbf{p} \circ \boldsymbol{\epsilon} = (p_N, \ldots, p_1)$. The composition $\boldsymbol{\epsilon} \circ \mathbf{p}$ results from the reversal of the order of the y values. So $\mathbf{p}(X, -Y) = \boldsymbol{\epsilon} \circ \mathbf{p}(X, Y)$. Similarly, the composition $\mathbf{p} \circ \boldsymbol{\epsilon}$ results from the reversal of the order of the x values, and so $\mathbf{p}(-X, Y) = \mathbf{p}(X, Y) \circ \boldsymbol{\epsilon}$. Now we shall motivate our definition of the correlation coefficient R_g.

* Rudy A. Gideon is Professor, Department of Mathematical Sciences, University of Montana, Missoula, MT 59812. Robert A. Hollister is Assistant Professor, Mathematics Department, University of Wisconsin, Oshkosh, WI 54901. Part of this work appears in Hollister's doctoral dissertation at the University of Montana. The authors would like to thank Michael J. Prentice, Edinburgh University, Scotland for his help on the population interpretation section while Gideon was on his sabbatical in Edinburgh. In addition, the authors appreciate referees' comments, which aided in the article's emphasis and in connecting the population interpretation to existing literature.

Journal of the American Statistical Association
June 1987, Vol. 82, No. 398, Theory and Methods

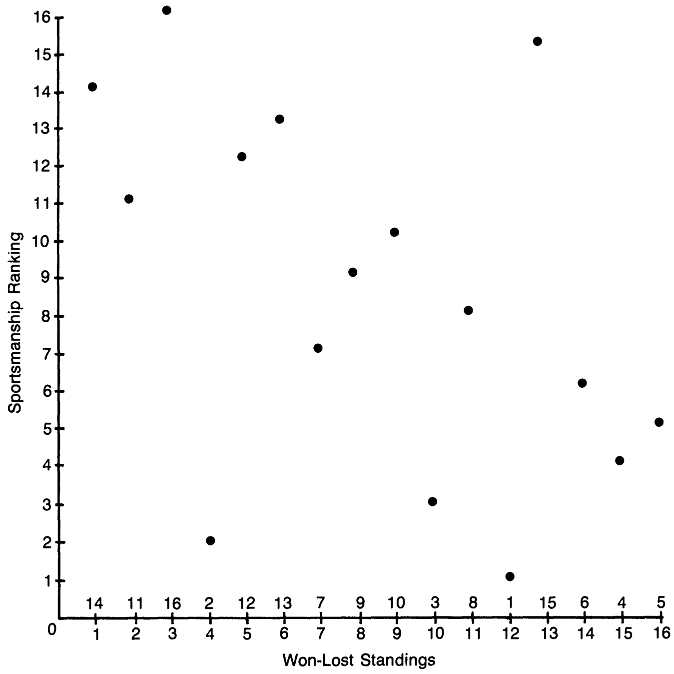

Figure 1. YMCA Basketball Data.

When the permutation for the data is the identity permutation $\mathbf{e}$ (reverse permutation $\boldsymbol{\epsilon}$), any rank correlation coefficient should be 1 (-1). Our new correlation coefficient is based on the property of maximum deviation of $\mathbf{p}(X, Y)$ from $\mathbf{e}$ and $\boldsymbol{\epsilon}$, that is, from permutations that represent perfect positive and negative correlation.

In comparing the permutation determined by the sample $\mathbf{p}(X, Y)$ with $\mathbf{e}$, we measure the deviation at i (for $i = 1, 2, \ldots, N$) by the number of $p_1, \ldots, p_i$ that exceed $e_i = i$.

Definition 1. Let $I(E) = 1$ if E is true and 0 if E is false, and let

$$d_i(\mathbf{p}) = \sum_{j=1}^{i} I(i < p_j) = \sum_{j=1}^{N} I(r(x_j) \le i < r(y_j)).$$

For the YMCA data, $(d_1(\mathbf{p}), d_2(\mathbf{p}), \ldots, d_{16}(\mathbf{p})) = (1, 2, 3, 3, 4, 5, 5, 6, 6, 5, 4, 3, 3, 2, 1, 0)$.

In comparing $\mathbf{p}(X, Y)$ with $\boldsymbol{\epsilon}$, we shall measure the deviation at i (for $i = 1, 2, \ldots, N$) by the number of $p_1, \ldots, p_i$ that are less than $\epsilon_i = N + 1 - i$. This is equivalent to measuring the deviation at i for $\boldsymbol{\epsilon} \circ \mathbf{p}$ with $\mathbf{e}$, since

$$\sum_{j=1}^{i} I(p_j < N + 1 - i) = \sum_{j=1}^{i} I(i < N + 1 - p_j) = d_i(\boldsymbol{\epsilon} \circ \mathbf{p}).$$

Again for the YMCA data, $\boldsymbol{\epsilon} \circ \mathbf{p} = (3, 6, 1, 15, 5, 4, 10, 8, 7, 14, 9, 16, 2, 11, 13, 12)$, and $(d_1(\boldsymbol{\epsilon} \circ \mathbf{p}), d_2(\boldsymbol{\epsilon} \circ \mathbf{p}), \ldots, d_{16}(\boldsymbol{\epsilon} \circ \mathbf{p})) = (1, 2, 1, 2, 2, 1, 2, 2, 2, 2, 2, 3, 3, 2, 1, 0)$.

Definition 2. $d(\mathbf{p}) = \max_i d_i(\mathbf{p})$. Then $d(\boldsymbol{\epsilon} \circ \mathbf{p}) = \max_i d_i(\boldsymbol{\epsilon} \circ \mathbf{p})$, and for the YMCA data $d(\mathbf{p}) = 6$ and $d(\boldsymbol{\epsilon} \circ \mathbf{p}) = 3$.

Definition 3. $R_g(X, Y) = (d(\boldsymbol{\epsilon} \circ \mathbf{p}) - d(\mathbf{p}))/[N/2]$, where $\mathbf{p} = \mathbf{p}(X, Y)$, the permutation determined by the sample, and $[\cdot]$ is the greatest integer notation. If we now compute R_g for the data of Figure 1, we have $R_g = (3 - 6)/[16/2] = -\frac{3}{8}$.

The statistic $d(\mathbf{p})(d(\boldsymbol{\epsilon} \circ \mathbf{p}))$ measures the greatest deviation of $\mathbf{p}$ from $\mathbf{e}$ ($\mathbf{p}$ from $\boldsymbol{\epsilon}$). The subscript g is used on R to denote greatest deviation. R_g is 1 if $\mathbf{p} = \mathbf{e}$, -1 if $\mathbf{p} = \boldsymbol{\epsilon}$, and 0 if $\mathbf{p}$ deviates from $\mathbf{e}$ and $\boldsymbol{\epsilon}$ equally.

3. PROPERTIES OF R_g

Reasonable correlation coefficients need to possess certain properties. Renyi (1959) gave a list of desirable properties for correlation coefficients and Schweizer and Wolfe (1981) gave a modified list of properties for nonparametric measures of dependence for continuously distributed random variables X and Y. This latter list is used to illustrate some of the properties that have been proved for R_g. In general the proofs are long and tedious and are, therefore, deleted, except for an outline of the proof of Property 3. The proofs appear in Hollister (1984).

Consider $R_g(X, Y)$ as a random variable, distributed over all possible samples of size N obtained from a continuous bivariate distribution of the random variables X and Y. Then the following properties hold.

Property 1. $R_g(X, Y)$ is well defined.
Property 2. $-1 \le R_g(X, Y) \le +1$.
Property 3. $R_g(Y, X) = R_g(X, Y)$.
Property 4. $R_g(-X, Y) = R_g(X, -Y) = -R_g(X, Y)$.

Table 1. YMCA Correlations and Probability Values

	(a) Teams 4 and 13 interchanged (more consistent)	Original data	(b) Teams 1 and 16 interchanged (less consistent)
R_g	$-\frac{1}{2} = -.500$	$-\frac{3}{8} = -.375$	$-\frac{1}{4} = -.250$
p value	.009	.068	.149
R_k	$-\frac{37}{60} = -.617$	$-\frac{11}{30} = -.367$	$-\frac{1}{12} = -.083$
p value	<.005	<.025	.326
R_s	$-\frac{283}{340} = -.832$	$-\frac{83}{170} = -.488$	$-\frac{31}{340} = -.091$
p value	<.001	.030	.362
R_q	$-\frac{3}{4} = -.750$	$-\frac{1}{2} = -.500$	$-\frac{1}{4} = -.250$
p value	.005	.066	.310

Property 5. $R_g(X, Y) = +1$ with probability 1 iff Y is a strictly monotone increasing function of X. $R_g(X, Y) = -1$ with probability 1 iff Y is a strictly monotone decreasing function of X.

Property 6. If X and Y are independent, then $E[R_g(X, Y)] = 0$.

Property 7. $R_g(f(X), g(Y)) = R_g(X, Y)$ if f and g are strictly monotone increasing functions on the ranges of X and Y, respectively.

In addition to these properties, several other facts about R_g have been proved, but again the proofs will be omitted. For the most part the proofs involved the properties of S_N and its operation $\circ$.

(a) For any positive integer N greater than 2, $[N/2]R_g(X, Y)$ can assume the $2[N/2] + 1$ values $k/[N/2]$ for $k = -[N/2], -[N/2] + 1, \ldots, -1, 0, 1, \ldots, [N/2]$.

(b) $P(R_g(X, Y) = +1) = P(R_g(X, Y) = -1) = 1/N!$, when X and Y are independent.

(c) The null distribution (X, Y independent) of $R_g(X, Y)$ is symmetric about 0.

(d) If $\mathbf{p} \circ \boldsymbol{\epsilon}$ replaces $\boldsymbol{\epsilon} \circ \mathbf{p}$ in the definition of R_g, then R_g remains unchanged, since it can be shown that $d(\mathbf{p} \circ \boldsymbol{\epsilon}) = d(\boldsymbol{\epsilon} \circ \mathbf{p})$. However, $d_i(\boldsymbol{\epsilon} \circ \mathbf{p}) = d_{N-i}(\mathbf{p} \circ \boldsymbol{\epsilon})$.

The technique used to prove these properties is illustrated by the following outline of our proof of Property 3.

Let $\mathbf{p}^{-1}$ be the inverse of $\mathbf{p}$. Then $\mathbf{p} \circ \mathbf{p}^{-1} = \mathbf{p}^{-1} \circ \mathbf{p} = \mathbf{e}$. Distinguish $\mathbf{p} = \mathbf{p}(X, Y)$ from $\mathbf{p}_y = \mathbf{p}(Y, X)$. Then $\mathbf{p}_y = \mathbf{p}^{-1}$. Thus

$$[N/2]R_g(Y, X) = d(\boldsymbol{\epsilon} \circ \mathbf{p}_y) - d(\mathbf{p}_y)$$

$$= d(\boldsymbol{\epsilon} \circ \mathbf{p}^{-1}) - d(\mathbf{p}^{-1})$$

$$= d((\mathbf{p} \circ \boldsymbol{\epsilon})^{-1}) - d(\mathbf{p}^{-1}), \text{ since } (\mathbf{p} \circ \boldsymbol{\epsilon})^{-1} = \boldsymbol{\epsilon}^{-1} \circ \mathbf{p}^{-1} = \boldsymbol{\epsilon} \circ \mathbf{p}^{-1};$$

$$= d(\mathbf{p} \circ \boldsymbol{\epsilon}) - d(\mathbf{p}), \text{ since } d(\mathbf{p}) = d(\mathbf{p}^{-1})$$

[because $d_i(\mathbf{p}) = d_i(\mathbf{p}^{-1})$, for all i].

Thus $[N/2]R_g(Y, X) = [N/2]R_g(X, Y)$ as $d(\mathbf{p} \circ \boldsymbol{\epsilon}) = d(\boldsymbol{\epsilon} \circ \mathbf{p})$ from Property (d).

4. THE DISTRIBUTION OF R_g AND SOME POWER COMPARISONS

The distribution of $R_g(X, Y)$ is directly related to that of $\mathbf{p}(X, Y)$, which is difficult to determine in most cases. Under the hypothesis of independence between X and Y (the null hypothesis for a test of independence), however, it becomes easier. In that case all of the permutations in S_N are equally likely. Thus $P(\mathbf{p}(X, Y) = \mathbf{p}) = 1/N!$ for each $\mathbf{p}$ in S_N.

The null distribution of R_g has been determined for sample sizes $N = 2$ to 10 by explicitly computing and tallying the value of R_g for every permutation in S_N with the aid of a computer. These distributions are tabulated in Table 2. For larger sample sizes (11–100), the distribution has been approximated using computer simulations. Two-sided randomized critical values for $\alpha = .01$, .05, .10 are listed in Table 3 (exact for $n \leq 10$ and approximations for $n > 10$). For sample sizes 100 to 500, Figure 2 is provided to allow interpolation for approximate critical values. Currently no explicit formula has been determined for the null distribution of R_g, nor has its asymptotic distribution been derived.

Table 2. The Null Distribution of R_g for $N = 2$ to 10 (symmetric about 0)

N	R_g	Frequency	Probability
2	1	1	.5000
3	1	1	.1667
	0	4	.6667
4	1	1	.0417
	$\frac{1}{2}$	3	.1250
	0	16	.6667
5	1	1	.0083
	$\frac{1}{2}$	51	.4250
	0	16	.1333
6	1	1	.0014
	$\frac{2}{3}$	35	.0486
	$\frac{1}{3}$	196	.2722
	0	256	.3556
7	1	1	.0002
	$\frac{2}{3}$	595	.1181
	$\frac{1}{3}$	500	.0992
	0	2848	.5651
8	1	1	.0000
	$\frac{3}{4}$	399	.0099
	$\frac{1}{2}$	2480	.0615
	$\frac{1}{4}$	11772	.2920
	0	11016	.2732
9	1	1	.0000
	$\frac{3}{4}$	6927	.0191
	$\frac{1}{2}$	18992	.0523
	$\frac{1}{4}$	123660	.3408
	0	63720	.1756
10	1	1	.0000
	$\frac{4}{5}$	4623	.0013
	$\frac{3}{5}$	36672	.0101
	$\frac{2}{5}$	479120	.1320
	$\frac{1}{5}$	562932	.1551
	0	1462104	.4029

To compare the power of R_g with other nonparametric correlation coefficients—Spearman's rho (R_s), Kendall's tau (R_k), and the quadrant correlation coefficient (R_q)—computer simulations were run for randomized two-sided tests of independence. For each sample size ($N = 5$, 6, 16, 20, 21, 25, 40) and level of significance ($\alpha = .01$, .05, .10) 10,000 random simulations were run. The samples were simulated from populations that were bivariate normal, bivariate exponential, and bivariate normal contaminated with biased outliers. The bivariate populations had correlations of $\rho = 0$, .3, .6, and .9. Of these comparisons, those selected for presentation exemplify the general conclusions deduced from all of the simulations.

Because of the discrete nature of the distribution of rank correlation coefficients, good power comparisons depend on using randomized tests to achieve the same α level for all compared statistics, and hence Table 3 is given to allow

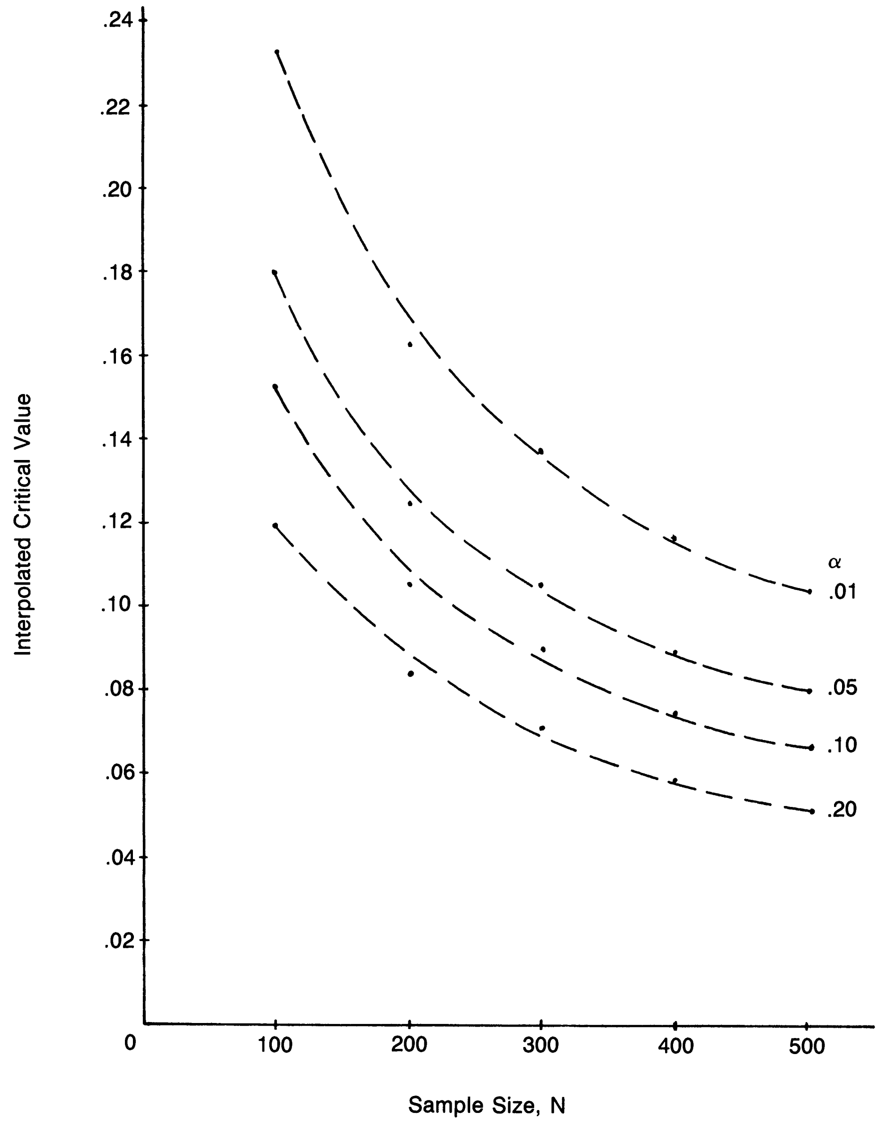

Figure 2. Interpolated Critical Values for a Randomized Two-Tailed Test of Independence, Using $R_g(X, Y)$, for Samples of Sizes 100, 200, 300, 400, and 500 (based on simulations of size 2,500). Interpolated critical values = $c_1 + \rho^*(c_1 - c_2)$, where $c_1 > c_2$ and $P(R_g \geq c_1) + \rho^* P(R_g = c_2) = \alpha$.

possible future comparisons. To use Table 3 for a two-sided test with $\alpha \leq .10$ for $N = 33$, reject the independence hypothesis if $|R_g| \geq \frac{5}{16}$; that is, use the column labeled Crit1. To get $\alpha = .10$ exactly, reject if $|R_g| \geq \frac{5}{16}$ and reject with probability $p = .4909$ if $|R_g| = \frac{4}{16}$.

The biased outliers referred to previously are based on the type of bias that may occur when comparing judges' rankings, for example, in diving or gymnastics competition. For instance, two judges from rival regions may each rank competitors from their own region more favorably than other competitors and rank those from the rival region more harshly. In the YMCA data, the YMCA director's son was on the fourth best team! In the simulations that were run, this bias was created by sampling from a bivariate normal distribution having the given correlation ρ and having standard normal marginal distributions. Then, if the first value of the pair sampled was an extreme value (e.g., absolute value of the sample exceeds $z_{\alpha/2}$), the second value was negated. For example, if $\alpha = .05$ and the pair sampled was (2, 1.5), then since $2 > z_{.025} = 1.96$, (2, 1.5) was replaced by the pair (2, −1.5) in the sample to test the hypothesis of independence. This is an exaggerated biased outlier concept and hence is useful for detecting effects of such outliers.

As expected, simulations from a bivariate normal population showed that the power of R_g was better than that of the quadrant correlation coefficient (R_q) and not as good as that of Spearman's rho (R_s) and Kendall's tau (R_k). Figure 3 illustrates this for $\rho = .6$. The power of the Pearson product moment correlation coefficient (R_p) was graphed for the same set of simulations.

When the sample was derived from a bivariate exponential population (Marshall and Olkin 1967), the power of R_g was better than that of the quadrant correlation (R_q) and not as good as Kendall's tau (R_k). Note, however, that the power of R_g was close to that of R_s. Moreover, the power of R_g overtook the power of the Spearman rho (R_s) as the sample size increased. For larger samples the

Table 3. Critical Values for a Randomized Two-Tailed Test of Independence Using $R_g(X, Y)$,[a] for Samples of Sizes 2 to 100[b]

	10%			5%			1%		
N	*Crit1*	*Crit2*	ρ	*Crit1*	*Crit2*	ρ	*Crit1*	*Crit2*	ρ
2	*****	$\frac{1}{1}$	.10000	*****	$\frac{1}{1}$	.05000	*****	$\frac{1}{1}$	.01000
3	*****	$\frac{1}{1}$	.30000	*****	$\frac{1}{1}$	.15000	*****	$\frac{1}{1}$	.03000
4	$\frac{2}{2}$	$\frac{1}{2}$	.06667	*****	$\frac{2}{2}$	.60000	*****	$\frac{2}{2}$	.12000
5	$\frac{2}{2}$	$\frac{1}{2}$	.09804	$\frac{2}{2}$	$\frac{1}{2}$	.03922	*****	$\frac{2}{2}$	.60000
6	$\frac{2}{3}$	$\frac{1}{3}$	.00000	$\frac{3}{3}$	$\frac{2}{3}$	.48571	$\frac{3}{3}$	$\frac{2}{3}$	.07429
7	$\frac{3}{3}$	$\frac{2}{3}$	.42185	$\frac{3}{3}$	$\frac{2}{3}$	.21008	$\frac{3}{3}$	$\frac{2}{3}$	.04067
8	$\frac{3}{4}$	$\frac{2}{4}$	.65161	$\frac{3}{4}$	$\frac{2}{4}$	.24516	$\frac{4}{4}$	$\frac{3}{4}$	.50276
9	$\frac{3}{4}$	$\frac{2}{4}$	.59056	$\frac{3}{4}$	$\frac{2}{4}$	.11289	$\frac{4}{4}$	$\frac{3}{4}$	.26179
10	$\frac{3}{5}$	$\frac{2}{5}$	.29250	$\frac{3}{5}$	$\frac{2}{5}$	.10316	$\frac{4}{5}$	$\frac{3}{5}$	.36867
11	$\frac{3}{5}$	$\frac{2}{5}$	.23640	$\frac{3}{5}$	$\frac{2}{5}$	.04260	$\frac{4}{5}$	$\frac{3}{5}$	.19350
12	$\frac{3}{6}$	$\frac{2}{6}$	.13820	$\frac{4}{6}$	$\frac{3}{6}$	.59640	$\frac{4}{6}$	$\frac{3}{6}$	.08190
13	$\frac{3}{6}$	$\frac{2}{6}$	.04960	$\frac{4}{6}$	$\frac{3}{6}$	.59530	$\frac{4}{6}$	$\frac{3}{6}$	.01480
14	$\frac{3}{7}$	$\frac{2}{7}$	.10300	$\frac{4}{7}$	$\frac{3}{7}$	.64540	$\frac{5}{7}$	$\frac{4}{7}$	.55230
15	$\frac{4}{7}$	$\frac{3}{7}$	.94240	$\frac{4}{7}$	$\frac{3}{7}$	.36640	$\frac{5}{7}$	$\frac{4}{7}$	.50000
16	$\frac{4}{8}$	$\frac{3}{8}$	.65860	$\frac{4}{8}$	$\frac{3}{8}$	.40770	$\frac{5}{8}$	$\frac{4}{8}$	.48750
17	$\frac{4}{8}$	$\frac{3}{8}$	.76300	$\frac{4}{8}$	$\frac{3}{8}$	.40380	$\frac{5}{8}$	$\frac{4}{8}$	.20000
18	$\frac{4}{9}$	$\frac{3}{9}$	.76600	$\frac{4}{9}$	$\frac{3}{9}$	.34650	$\frac{5}{9}$	$\frac{4}{9}$	.17750
19	$\frac{4}{9}$	$\frac{3}{9}$	.50390	$\frac{4}{9}$	$\frac{3}{9}$	.34400	$\frac{5}{9}$	$\frac{4}{9}$	.06110
20	$\frac{4}{10}$	$\frac{3}{10}$	.41240	$\frac{4}{10}$	$\frac{3}{10}$	.05320	$\frac{5}{10}$	$\frac{4}{10}$	.01510
21	$\frac{4}{10}$	$\frac{3}{10}$	.40810	$\frac{5}{10}$	$\frac{4}{10}$	.81180	$\frac{5}{10}$	$\frac{4}{10}$	.02740
22	$\frac{4}{11}$	$\frac{3}{11}$	.29500	$\frac{5}{11}$	$\frac{4}{11}$	.43330	$\frac{6}{11}$	$\frac{5}{11}$	.58820
23	$\frac{4}{11}$	$\frac{3}{11}$	.27790	$\frac{5}{11}$	$\frac{4}{11}$	.43150	$\frac{6}{11}$	$\frac{5}{11}$	.47060
24	$\frac{4}{12}$	$\frac{3}{12}$	.18570	$\frac{5}{12}$	$\frac{4}{12}$	.38160	$\frac{6}{12}$	$\frac{5}{12}$	.34440
25	$\frac{4}{12}$	$\frac{3}{12}$	.06400	$\frac{5}{12}$	$\frac{4}{12}$	.36730	$\frac{6}{12}$	$\frac{5}{12}$	.27400
26	$\frac{4}{13}$	$\frac{3}{13}$	.07220	$\frac{5}{13}$	$\frac{4}{13}$	.33850	$\frac{6}{13}$	$\frac{5}{13}$	.05260
27	$\frac{4}{13}$	$\frac{3}{13}$	.10510	$\frac{5}{13}$	$\frac{4}{13}$	.32930	$\frac{6}{13}$	$\frac{5}{13}$	.16410
28	$\frac{5}{14}$	$\frac{4}{14}$	.95360	$\frac{5}{14}$	$\frac{4}{14}$	.22900	$\frac{6}{14}$	$\frac{5}{14}$	.06340
29	$\frac{5}{14}$	$\frac{4}{14}$	.81450	$\frac{5}{14}$	$\frac{4}{14}$	.18800	$\frac{7}{14}$	$\frac{6}{14}$	.92000
30	$\frac{5}{15}$	$\frac{4}{15}$	.83330	$\frac{5}{15}$	$\frac{4}{15}$	.09480	$\frac{7}{15}$	$\frac{6}{15}$	.75000
31	$\frac{5}{15}$	$\frac{4}{15}$	.79740	$\frac{5}{15}$	$\frac{4}{15}$	.00000	$\frac{7}{15}$	$\frac{6}{15}$	.63460
32	$\frac{5}{16}$	$\frac{4}{16}$	.63730	$\frac{5}{16}$	$\frac{4}{16}$	.01660	$\frac{7}{16}$	$\frac{6}{16}$	.67920
33	$\frac{5}{16}$	$\frac{4}{16}$	.49090	$\frac{6}{16}$	$\frac{5}{16}$	.92900	$\frac{7}{16}$	$\frac{6}{16}$	.53130
34	$\frac{5}{17}$	$\frac{4}{17}$	.50380	$\frac{6}{17}$	$\frac{5}{17}$	.62740	$\frac{7}{17}$	$\frac{6}{17}$	.47830
35	$\frac{5}{17}$	$\frac{4}{17}$	.40050	$\frac{6}{17}$	$\frac{5}{17}$	.67940	$\frac{7}{17}$	$\frac{6}{17}$	.29730
36	$\frac{5}{18}$	$\frac{4}{18}$	.35230	$\frac{6}{18}$	$\frac{5}{18}$	.61810	$\frac{7}{18}$	$\frac{6}{18}$	.20730
37	$\frac{5}{18}$	$\frac{4}{18}$	.33620	$\frac{6}{18}$	$\frac{5}{18}$	.54400	$\frac{7}{18}$	$\frac{6}{18}$	.10620
38	$\frac{5}{19}$	$\frac{4}{19}$	.26040	$\frac{6}{19}$	$\frac{5}{19}$	.36090	$\frac{7}{19}$	$\frac{6}{19}$	.02080
39	$\frac{5}{19}$	$\frac{4}{19}$	.26320	$\frac{6}{19}$	$\frac{5}{19}$	.35360	$\frac{7}{19}$	$\frac{6}{19}$	.02220
40	$\frac{5}{20}$	$\frac{4}{20}$	.23800	$\frac{6}{20}$	$\frac{5}{20}$	.29910	$\frac{7}{20}$	$\frac{6}{20}$	.05940
41	$\frac{5}{20}$	$\frac{4}{20}$	.14670	$\frac{6}{20}$	$\frac{5}{20}$	.31780	$\frac{8}{20}$	$\frac{7}{20}$	.87100
42	$\frac{5}{21}$	$\frac{4}{21}$	.00810	$\frac{6}{21}$	$\frac{5}{21}$	.31090	$\frac{8}{21}$	$\frac{7}{21}$	.76320
43	$\frac{6}{21}$	$\frac{5}{21}$	.94700	$\frac{6}{21}$	$\frac{5}{21}$	.16480	$\frac{8}{21}$	$\frac{7}{21}$	.40910
44	$\frac{6}{22}$	$\frac{5}{22}$	.84330	$\frac{6}{22}$	$\frac{5}{22}$	.14290	$\frac{8}{22}$	$\frac{7}{22}$	.37290
45	$\frac{6}{22}$	$\frac{5}{22}$	.86220	$\frac{7}{22}$	$\frac{6}{22}$	.94510	$\frac{8}{22}$	$\frac{7}{22}$	.34380
46	$\frac{6}{23}$	$\frac{5}{23}$	.78920	$\frac{6}{23}$	$\frac{5}{23}$	.08310	$\frac{8}{23}$	$\frac{7}{23}$	.30190
47	$\frac{6}{23}$	$\frac{5}{23}$	.74230	$\frac{7}{23}$	$\frac{6}{23}$	.88270	$\frac{8}{23}$	$\frac{7}{23}$	.10840
48	$\frac{6}{24}$	$\frac{5}{24}$	.61240	$\frac{7}{24}$	$\frac{6}{24}$	.98560	$\frac{8}{24}$	$\frac{7}{24}$	.09590
49	$\frac{6}{24}$	$\frac{5}{24}$	.71890	$\frac{7}{24}$	$\frac{6}{24}$	.79680	$\frac{8}{24}$	$\frac{7}{24}$	.09520
50	$\frac{6}{25}$	$\frac{5}{25}$	.55620	$\frac{7}{25}$	$\frac{6}{25}$	.65660	$\frac{8}{25}$	$\frac{7}{25}$	.17390
51	$\frac{6}{25}$	$\frac{5}{25}$	.43660	$\frac{7}{25}$	$\frac{6}{25}$	.58060	$\frac{9}{25}$	$\frac{8}{25}$	.75760
52	$\frac{6}{26}$	$\frac{5}{26}$	.58200	$\frac{7}{26}$	$\frac{6}{26}$	.55630	$\frac{9}{26}$	$\frac{8}{26}$	.91180

powers of R_g and R_s were essentially equal, with that of R_g generally being slightly greater. Figure 4 illustrates this for $\rho = .6$. Again the power of the Pearson product moment correlation coefficient (R_p) was included for comparison even though it is not appropriate for this distribution.

When the samples were bivariate normal with the biased outlier contamination, the powers of the correlation coefficients were ordered as they were for the pure bivariate normal case when the sample was quite small. However, the power of R_g increased relative to the others as the sample size increased. R_g had the most power for larger samples. Figure 5 illustrates this for simulations from a bivariate normal with $\rho = .6$, which was contaminated by biased outliers as explained earlier.

Further study of biased outliers showed that Spearman's rho (R_s) and Kendall's tau (R_k) often rejected the null hypothesis of independence in the wrong direction, whereas R_g rarely did. That is, when $\rho > 0$, the rejection was frequently due to the sample correlation being more *negative* than the negative critical value. In this case we shall say that the null hypothesis was incorrectly rejected. The Pearson product moment correlation coefficient (R_p) is extremely sensitive to this contamination. Table 4 gives

Table 3 (continued)

	10%			5%			1%		
N	Crit1	Crit2	p	Crit1	Crit2	p	Crit1	Crit2	p
53	$\frac{6}{26}$	$\frac{5}{26}$	.51160	$\frac{7}{26}$	$\frac{6}{26}$	.57320	$\frac{9}{26}$	$\frac{8}{26}$	.55260
54	$\frac{6}{27}$	$\frac{5}{27}$	.26920	$\frac{7}{27}$	$\frac{6}{27}$	.42960	$\frac{9}{27}$	$\frac{8}{27}$	.57690
55	$\frac{6}{27}$	$\frac{5}{27}$	.32860	$\frac{7}{27}$	$\frac{6}{27}$	.49870	$\frac{9}{27}$	$\frac{8}{27}$	.75860
56	$\frac{6}{28}$	$\frac{5}{28}$	.24570	$\frac{7}{28}$	$\frac{6}{28}$	.13710	$\frac{9}{28}$	$\frac{8}{28}$	.55000
57	$\frac{6}{28}$	$\frac{5}{28}$	.30410	$\frac{7}{28}$	$\frac{6}{28}$	.31330	$\frac{9}{28}$	$\frac{8}{28}$	.33330
58	$\frac{6}{29}$	$\frac{5}{29}$	.12240	$\frac{7}{29}$	$\frac{6}{29}$	.31290	$\frac{9}{29}$	$\frac{8}{29}$	.16670
59	$\frac{6}{29}$	$\frac{5}{29}$	.06640	$\frac{7}{29}$	$\frac{6}{29}$	.13770	$\frac{9}{29}$	$\frac{8}{29}$	.18000
60	$\frac{6}{30}$	$\frac{5}{30}$	.07990	$\frac{7}{30}$	$\frac{6}{30}$	.13490	$\frac{9}{30}$	$\frac{8}{30}$	.05560
61	$\frac{6}{30}$	$\frac{5}{30}$	.11670	$\frac{7}{30}$	$\frac{6}{30}$	.00550	$\frac{9}{30}$	$\frac{8}{30}$	.05450
62	$\frac{6}{31}$	$\frac{5}{31}$	.01990	$\frac{7}{31}$	$\frac{7}{31}$	.95240	$\frac{9}{31}$	$\frac{8}{31}$	.04240
63	$\frac{6}{31}$	$\frac{5}{31}$	.04650	$\frac{7}{31}$	$\frac{6}{31}$	.05740	$\frac{9}{31}$	$\frac{8}{31}$	.06370
64	$\frac{7}{32}$	$\frac{6}{32}$	.89510	$\frac{8}{32}$	$\frac{7}{32}$	.91180	$\frac{10}{32}$	$\frac{9}{32}$	.87440
65	$\frac{7}{32}$	$\frac{6}{32}$	.86900	$\frac{8}{32}$	$\frac{7}{32}$	.78620	$\frac{10}{32}$	$\frac{9}{32}$	.71920
66	$\frac{7}{33}$	$\frac{6}{33}$	.84880	$\frac{8}{33}$	$\frac{7}{33}$	.81300	$\frac{10}{33}$	$\frac{9}{33}$	.79170
67	$\frac{7}{33}$	$\frac{6}{33}$	.63290	$\frac{8}{33}$	$\frac{7}{33}$	.77600	$\frac{10}{33}$	$\frac{9}{33}$	.58620
68	$\frac{7}{34}$	$\frac{6}{34}$	.71190	$\frac{8}{34}$	$\frac{7}{34}$	.75000	$\frac{10}{34}$	$\frac{9}{34}$	.75000
69	$\frac{7}{34}$	$\frac{6}{34}$	.66500	$\frac{8}{34}$	$\frac{7}{34}$	.61310	$\frac{10}{34}$	$\frac{9}{34}$	.57690
70	$\frac{7}{35}$	$\frac{6}{35}$	.82900	$\frac{8}{35}$	$\frac{7}{35}$	.39040	$\frac{10}{35}$	$\frac{9}{35}$	.30560
71	$\frac{7}{35}$	$\frac{6}{35}$	.53970	$\frac{8}{35}$	$\frac{7}{35}$	.47370	$\frac{10}{35}$	$\frac{9}{35}$	.36110
72	$\frac{7}{36}$	$\frac{6}{36}$	.58200	$\frac{8}{36}$	$\frac{7}{36}$	.33820	$\frac{10}{36}$	$\frac{9}{36}$	.25000
73	$\frac{7}{36}$	$\frac{6}{36}$	.41960	$\frac{8}{36}$	$\frac{7}{36}$	.25640	$\frac{10}{36}$	$\frac{9}{36}$	.10260
74	$\frac{7}{37}$	$\frac{6}{37}$	.51880	$\frac{8}{37}$	$\frac{7}{37}$	.39390	$\frac{10}{37}$	$\frac{9}{37}$	.13640
75	$\frac{7}{37}$	$\frac{6}{37}$	.37900	$\frac{8}{37}$	$\frac{7}{37}$	.30720	$\frac{10}{37}$	$\frac{9}{37}$	.25580
76	$\frac{7}{38}$	$\frac{6}{38}$	.32220	$\frac{8}{38}$	$\frac{7}{38}$	.14570	$\frac{10}{38}$	$\frac{9}{38}$	.08700
77	$\frac{7}{38}$	$\frac{6}{38}$	.31350	$\frac{8}{38}$	$\frac{7}{38}$	.35250	$\frac{10}{38}$	$\frac{9}{38}$	.31580
78	$\frac{7}{39}$	$\frac{6}{39}$	.31490	$\frac{8}{39}$	$\frac{7}{39}$	.14570	$\frac{10}{39}$	$\frac{9}{39}$	.04000
79	$\frac{7}{39}$	$\frac{6}{39}$	.22090	$\frac{8}{39}$	$\frac{7}{39}$	.01430	$\frac{11}{39}$	$\frac{10}{39}$	.64710
80	$\frac{7}{40}$	$\frac{6}{40}$	.22050	$\frac{8}{40}$	$\frac{7}{40}$	.03390	$\frac{11}{40}$	$\frac{10}{40}$	.60000
81	$\frac{7}{40}$	$\frac{6}{40}$	.21190	$\frac{8}{40}$	$\frac{7}{40}$	.05260	$\frac{11}{40}$	$\frac{10}{40}$	.16670
82	$\frac{7}{41}$	$\frac{6}{41}$	.14910	$\frac{9}{41}$	$\frac{8}{41}$	.75440	$\frac{11}{41}$	$\frac{10}{41}$	.92310
83	$\frac{7}{41}$	$\frac{6}{41}$	.09230	$\frac{9}{41}$	$\frac{8}{41}$	.92310	$\frac{11}{41}$	$\frac{10}{41}$	.16670
84	$\frac{7}{42}$	$\frac{6}{42}$	.08940	$\frac{9}{42}$	$\frac{8}{42}$	.86000	$\frac{11}{42}$	$\frac{10}{42}$	.50000
85	$\frac{7}{42}$	$\frac{6}{42}$	.02960	$\frac{9}{42}$	$\frac{8}{42}$	.78460	$\frac{10}{42}$	$\frac{9}{42}$	.06900
86	$\frac{8}{43}$	$\frac{7}{43}$	.86460	$\frac{9}{43}$	$\frac{8}{43}$	.88890	$\frac{11}{43}$	$\frac{10}{43}$	.38460
87	$\frac{8}{43}$	$\frac{7}{43}$	.95450	$\frac{9}{43}$	$\frac{8}{43}$	.64290	$\frac{10}{43}$	$\frac{9}{43}$	.03330
88	$\frac{7}{44}$	$\frac{6}{44}$	.04170	$\frac{9}{44}$	$\frac{8}{44}$	.79590	$\frac{11}{44}$	$\frac{10}{44}$	.00000
89	$\frac{7}{44}$	$\frac{6}{44}$	.05000	$\frac{9}{44}$	$\frac{8}{44}$	.77080	$\frac{12}{44}$	$\frac{11}{44}$	.91670
90	$\frac{8}{45}$	$\frac{7}{45}$	.69880	$\frac{9}{45}$	$\frac{8}{45}$	.61710	$\frac{12}{45}$	$\frac{11}{45}$	.85710
91	$\frac{8}{45}$	$\frac{7}{45}$	.55170	$\frac{9}{45}$	$\frac{8}{45}$	.43860	$\frac{11}{45}$	$\frac{10}{45}$	.33330
92	$\frac{8}{46}$	$\frac{7}{46}$	.55790	$\frac{9}{46}$	$\frac{8}{46}$	.37500	$\frac{11}{46}$	$\frac{10}{46}$	.00000
93	$\frac{8}{46}$	$\frac{7}{46}$	.64710	$\frac{9}{46}$	$\frac{8}{46}$	.32260	$\frac{11}{46}$	$\frac{10}{46}$	.28570
94	$\frac{8}{47}$	$\frac{7}{47}$	.63640	$\frac{9}{47}$	$\frac{8}{47}$	.48280	$\frac{11}{47}$	$\frac{10}{47}$	.12500
95	$\frac{8}{47}$	$\frac{7}{47}$	.63830	$\frac{9}{47}$	$\frac{8}{47}$	.39240	$\frac{12}{47}$	$\frac{11}{47}$	.27270
96	$\frac{8}{48}$	$\frac{7}{48}$	.51550	$\frac{9}{48}$	$\frac{8}{48}$	.16070	$\frac{11}{48}$	$\frac{10}{48}$	.20000
97	$\frac{8}{48}$	$\frac{7}{48}$	.55240	$\frac{9}{48}$	$\frac{8}{48}$	.28570	$\frac{12}{48}$	$\frac{11}{48}$	.25000
98	$\frac{8}{49}$	$\frac{7}{49}$	.59520	$\frac{9}{49}$	$\frac{8}{49}$	.13640	$\frac{12}{49}$	$\frac{11}{49}$	.52940
99	$\frac{8}{49}$	$\frac{7}{49}$	.61860	$\frac{9}{49}$	$\frac{8}{49}$	.07790	$\frac{12}{49}$	$\frac{11}{49}$	.68750
100	$\frac{8}{50}$	$\frac{7}{50}$	.35710	$\frac{9}{50}$	$\frac{8}{50}$	.00000	$\frac{12}{50}$	$\frac{11}{50}$	.72730

[a] Example: To obtain a two-sided test with $\alpha = .05$ for $N = 15$, reject H_0 : independent variables if $|R_g| \geq \frac{4}{7}$ and reject H_0 with probability $\rho = .36640$ if $|R_g| = \frac{3}{7}$.

[b] The values are based on the exact distribution of $R_g(X, Y)$ for $N = 2$ to 10 and on simulations (of size 10,000) for $N = 11$ to 100; thus the fifth decimal place (0) for $N > 11$ appears only as a visual convenience.

the results from 1,000 simulations of samples of the stated size from a bivariate normal population with the indicated correlation (ρ) and with biased outlier contamination. Consequently, for one-sided alternatives the power of R_g would be better relative to that of the other sample correlations.

5. TIED RANKS

A summary of tied rank procedures appears in Hájek and Šidak (1967, pp. 118–123), and we will assume that the reader is familiar with the randomization technique. For many data sets with tied values, R_g assumes only one value, and hence we recommend the randomization method so that Tables 2 and 3 can be used. We also recommend that the highest and lowest values of R_g be computed over the range of possible randomizations. If it is found that the difference between these values is large, then the conclusion should be drawn that there is little information in the data set.

To determine the extreme values of R_g, the two randomizations of **p** that most favor positive and negative correlation are determined.

Let us demonstrate the suggested procedure by taking an example from Conover (1980, example 1, p. 253). This

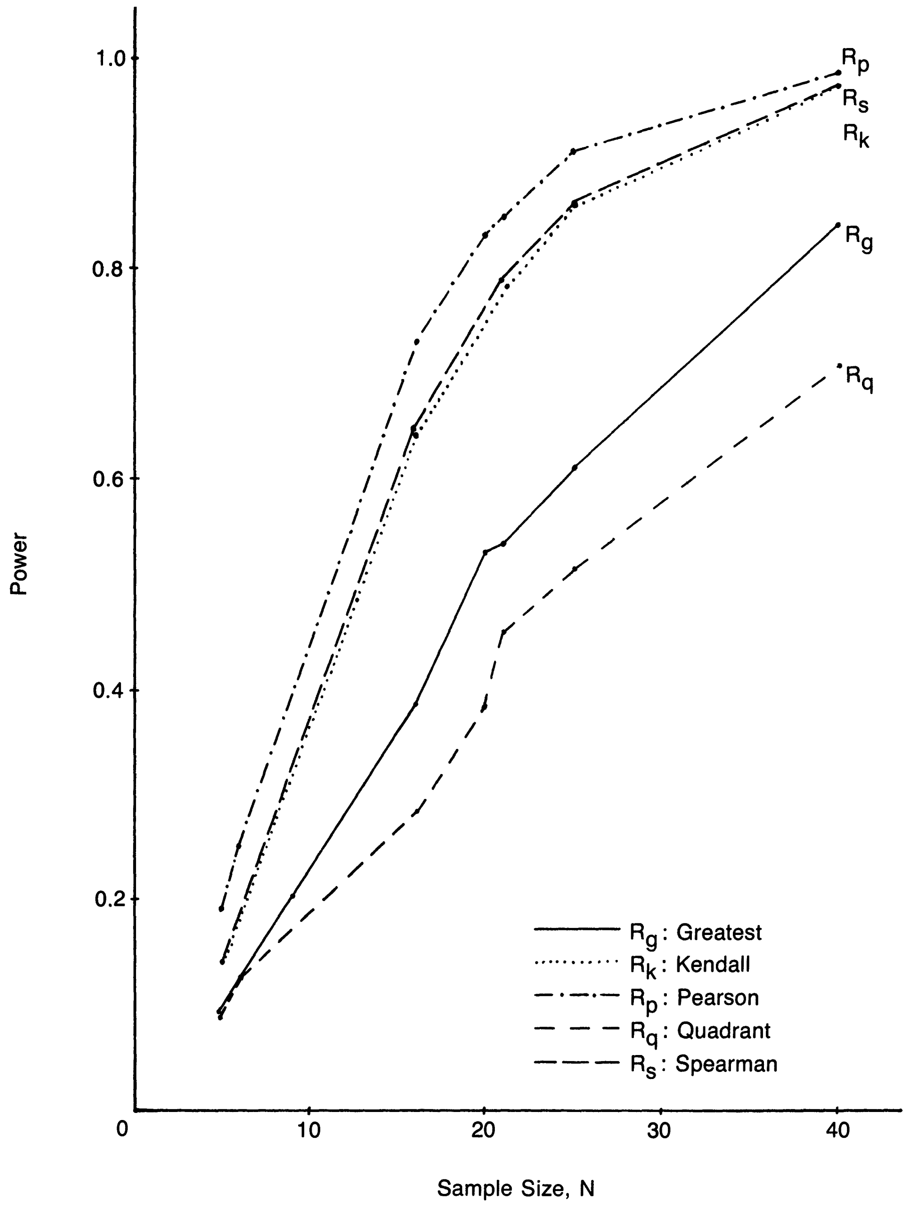

Figure 3. Relative Powers of Randomized Tests of Independence From a Bivariate Normal Population Based on 10,000 Simulations for Each of N = 5, 6, 16, 20, 21, 25, and 40 (ρ = .6, α = .05, two-tailed tests).

example is chosen because the tied rank procedure is discussed there for R_s and R_k. The data are from psychological tests on identical twins, with X being the first born. The data given are in the well-known mid-rank form:

X	1	2	3.5	3.5	5	6.5	6.5	8	9	10	11.5	11.5
Y	1	2.5	8	7	4.5	6	2.5	10	4.5	9	12	11

From Conover, R_s = .7378 or .7354 depending on which formula is used for R_s, and R_k = .5606. The approximate probability values for the two-sided test are given as .01 for both R_s and R_k.

To obtain the randomization of this tied data that most favors positive correlation, one simply chooses the lowest possible rank for Y as one proceeds over the 12 ranks of X within the constraints of the tied values. The permutation obtained is the same if the roles of X and Y have been interchanged. In a similar manner the permutation most favoring negative correlation is determined.

We list these two permutations:

X	1	2	3	4	5	6	7	8	9	10	11	12
Y (+ correlation)	1	2	7	8	4	3	6	10	5	9	11	12
Y (− correlation)	1	3	8	7	5	6	2	10	4	9	12	11

In both cases $R_g = (4 - 2)/6 = \frac{1}{3}$ and hence all randomizations would give $R_g = \frac{1}{3}$. For $N = 12$, R_g is significant at the 10% level, significant with probability .5964 at the 5% level, and significant with probability .0819 at the 1% level.

Thus, for this data set, the use of R_g leads to 10% significance, whereas R_s and R_k are approximate tests significant at the 1% level but based on limiting distributions

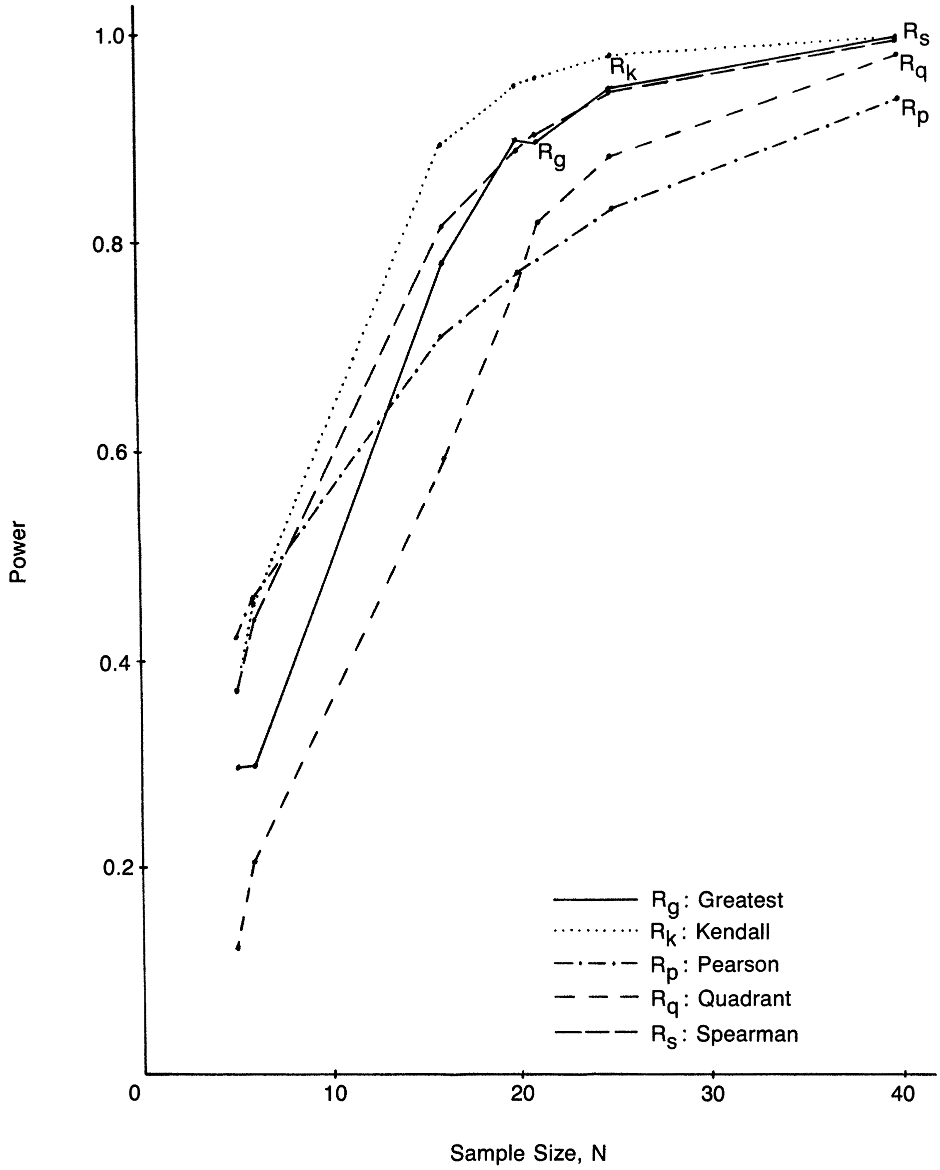

Figure 4. Relative Powers of Randomized Tests of Independence From a Bivariate Exponential Population Based on 10,000 Simulations for Each of N = 5, 6, 16, 20, 21, 25, and 40 (ρ = .6, α = .05, two-tailed tests).

that may be unreliable for small sample sizes. A possible use of R_g as an exact test for data sets with tied values would certainly help an experimenter in evaluating his data, especially when the sample size is very modest, as in this example.

The above data set was one of several that were checked in various nonparametric statistics books. Most led to one value of R_g. Suppose, however, that the X variable had all distinct ranks but all N Y ranks were tied. Then the rejection of the null hypothesis would be unrelated to the gathered data but would be entirely due to the randomization procedure. In this case the two extremes of R_g would be in -1 and $+1$ and all experimenters would realize that there is no information in their data relating X and Y. Note also that in this case, the average of the two extreme possible values of R_g would be 0.

The use of mid-ranks is well established for many rank statistics, but after some study, no satisfactory way was found for their use with R_g. On the other hand, the idea of determining the highest and lowest statistic over the range of possible permutations within the constraints of tied data might be beneficial descriptive statistics for other statistics besides R_g.

6. POPULATION INTERPRETATION OF R_g

Kruskal (1958) gave a population interpretation to R_s, R_k, and R_q. It is possible also to relate the correlation statistic R_g to a population parameter. Assume that a bivariate random variable (X, Y) is absolutely continuous and that a sample of size N is to be drawn. Let $X_{(i)}$, $Y_{(i)}$ be the order statistics. It is straightforward to show that for $R_g(X, Y)$ the quantity $d_i(\mathbf{p})/i$ equals, within the sample, the proportion of cases in which $Y > Y_{(i)}$ given $X \leq$

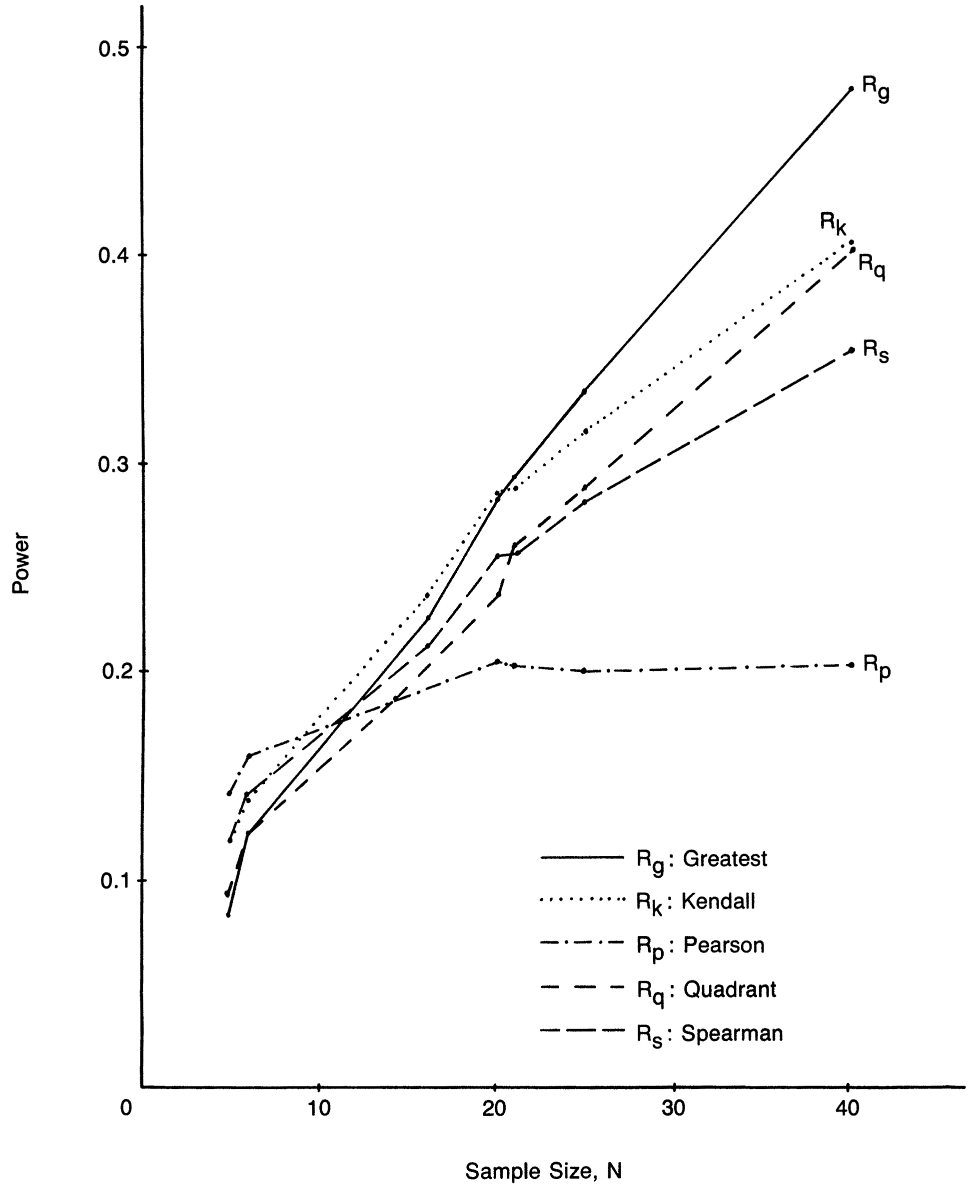

Figure 5. Relative Powers of Randomized Tests of Independence From a Bivariate Normal Population With 10% Biased Outliers Based on 10,000 Simulations for Each of N = 5, 6, 16, 20, 21, 25, and 40 (ρ = .6, α = .05, one-tailed tests).

$X_{(i)}$. Likewise, $d_i(\epsilon \circ \mathbf{p})/i$ equals the proportion of cases in which $Y < Y_{(N+1-i)}$ given $X \le X_{(i)}$. Thus $d_i(\mathbf{p})/i$ is an estimate of $P(Y > Y_{(i)} \mid X \le X_{(i)})$ and $d_i(\epsilon \circ \mathbf{p})/i$ is an estimate of $P(Y < Y_{(N+1-i)} \mid X \le X_{(i)})$. For $i = 1, 2, \ldots, N$, $P(Y > Y_{(i)} \mid X \le X_{(i)})$ is the standardized area of a series of rectangles [corner at $(X_{(i)}, Y_{(i)})$], which are open toward the upper left if we let X be the abscissa and Y be the ordinate axes. Similarly, as $i = 1, 2, \ldots, N$ $P(Y < Y_{(N+1-i)} \mid X \le X_{(i)})$ is the standardized area of a series of rectangles [corner at $(X_{(i)}, Y_{(N+1-i)})$], which are open toward the lower left.

To simplify matters let us use the probability integral transformation as was done in Kruskal (1958). Let $U = F(X)$ and $V = G(Y)$, where F and G are the marginal cdf's of X and Y, respectively. Then for the joint density of (U, V), the marginals will be $U(0, 1)$. Let $U_{(i)}$, $V_{(i)}$ be the order statistics for random variables U, V; (U, V) are called the grades in Kruskal. Then $d_i(\mathbf{p})/[N/2]$ will estimate

$$(i/[N/2])P(U \le U_{(i)}, V > V_{(i)})/P(U \le U_{(i)}),$$

and $d_i(\epsilon \circ \mathbf{p})/[N/2]$ will estimate

$$(i/[N/2])P(U \le U_{(i)}, V < V_{(N+1-i)})/P(U \le U_{(i)}).$$

For N large, $U_{(i)}$ approaches its expectation $i/(N + 1)$, and since $P(U \le i/(N + 1)) = i/(N + 1)$ and $([N/2]/i) \cdot (i/(N + 1)$ approaches $\frac{1}{2}$, for large N and letting $i/(N + 1) \to t$,

$$R_g = \max_i d_i(\epsilon \circ \mathbf{p})/[N/2] - \max_i d_i(\mathbf{p})/[N/2]$$

estimates

$$\sup_{0<t<1} 2P(U \le t, V < 1 - t) - \sup_{0<t<1} 2P(U \le t, V > t).$$

Table 4. Wrong Direction Rejection Comparisons for Biased Outlier Simulations

Correlation coefficient	Total number rejected	Number incorrectly rejected
Sample size = 20, ρ = .2, 1,000 samples		
R_g	34	7
R_k	56	32
R_s	55	30
R_p	57	40
Sample size = 21, ρ = .8, 1,000 samples		
R_g	138	2
R_k	140	49
R_s	113	57
R_p	263	229

Before proceeding with examples, let us relate the previous formula to the copula function C used in Schweizer and Wolfe (1981).

$$P(U \le t, V < 1 - t) = C(t, 1 - t),$$

and

$$P(U \le t, V > t) = C(t, 1) - C(t, t).$$

Thus in the limit

$$R_g = 2 \sup_{0<t<1} C(t, 1 - t) - 2 \sup_{0<t<1} [C(t, 1) - C(t, t)].$$

Now as stated in Schweizer and Wolfe (1981),

$$\begin{aligned} C(u, v) &= \max(u + v - 1, 0) \text{ for perfect negative correlation} \\ &= uv \text{ if independent} \\ &= \min(u, v) \text{ for perfect positive correlation.} \end{aligned}$$

Thus $C(t, 1 - t) = 0$ for perfect positive correlation and hence $\sup C(t, 1 - t) - 0$ measures the distance from perfect positive correlation. Likewise, $C(t, 1) - C(t, t) = 0$ for perfect negative correlation and $\sup[C(t, 1) - C(t, t)] - 0$ measures the distance from perfect negative correlation. The quantity $\kappa(X, Y) = 4 \sup_{0<u,v<1} |C(u, v) - uv|$ was introduced by Blum, Kiefer, and Rosenblatt (1961) as a test of independence, but it was not developed for practical use and its asymptotic distribution was not derived. In contrast to R_g their statistic measures distance from independence, and the sample statistic form

$$\hat{\kappa} = 4 \sup_{x,y} [H_n(x, y) - F_n(x)G_n(y)],$$

where H_n, F_n, G_n are empirical distribution functions, needs a computer for evaluation even for small sample sizes.

We now give two examples to show that R_g can sometimes behave like Kendall's tau and sometimes like Spearman's rho. If (X, Y) is bivariate normal, say

$$N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\rho & 1\\1 & \rho\end{pmatrix}\right),$$

then for large N, R_g estimates the same quantities as Kendall's τ, $(2/\pi)\sin^{-1}\rho$. To see this, we make the probability integral transformation and then $C(u, v)$, the copula function, is the bivariate cdf of (U, V). Then for large N,

$$\begin{aligned} R_g &= \sup_{0<t<1} 2C(t, 1 - t) - \sup_{0<t<1} 2(t - C(t, t)) \\ &= 2C\left(\frac{1}{2}, \frac{1}{2}\right) - 2\left(\frac{1}{2} - C\left(\frac{1}{2}, \frac{1}{2}\right)\right) \\ &= 2\left(\frac{1}{4} + \frac{1}{2\pi}\sin^{-1}\rho\right) - 2\left(\frac{1}{4} - \frac{1}{2\pi}\sin^{-1}\rho\right) \\ &= \frac{2}{\pi}\sin^{-1}\rho, \end{aligned}$$

because the bivariate normal has maximum probability of open rectangles at the medians. If $\rho = \frac{3}{4}$, then $R_g = R_k = (2/\pi)\sin^{-1}\frac{3}{4} = .5399$ and $R_s = (6/\pi)\sin^{-1}(\rho/2) = .7341$.

It is not true that R_g always estimates the same quantity that R_k does. To show this, take the following example, where the density of U, V is

$$\begin{aligned} g(u, v) &= 2 \text{ for } 0 \le u, v \le \tfrac{1}{2} \text{ and for } \tfrac{1}{2} \le u, v \le 1 \\ &= 0 \text{ elsewhere.} \end{aligned}$$

Then the marginals are $U(0, 1)$ and it is straightforward to show that $\rho = \frac{3}{4}$, $R_s = R_g = \frac{3}{4}$, but $R_k = \frac{1}{2}$.

Finally, if X and Y are independent, then so are U and V. In this case $\max_i d_i(\mathbf{p})/[N/2]$ and $\max_i d_i(\boldsymbol{\epsilon} \circ \mathbf{p})/[N/2]$ both estimate $\frac{1}{2}$ and R_g estimates $\sup_{0<t<1} 2t(1 - t) - \sup_{0<t<1} 2(t - t^2) = 0$.

7. FINAL COMMENTS

It should be noted that in the biased outlier simulations the quadrant correlation coefficient (R_q) also increased in power relative to R_s and R_k, becoming second to R_g for large samples. R_q is closely related to a correlation coefficient defined similarly to R_g but based on the deviation at only one point instead of the maximum deviations at all points. All results are stated without proofs, which are tedious but straightforward (Hollister 1984). To see this, define, for an integer $0 < i < N$, $R_i(X, Y)$ as follows: $R_i(X, Y) = (d_i(\boldsymbol{\epsilon} \circ \mathbf{p}) - d_i(\mathbf{p}))/N_i$, where $\mathbf{p} = \mathbf{p}(X, Y)$ and $N_i = \min(i, N - i)$. Under the null hypothesis of independence between X and Y, $d_i(\mathbf{p})$ and $d_i(\boldsymbol{\epsilon} \circ \mathbf{p})$ are hypergeometric random variables and $R_i(X, Y)$ has the probability function

$$\begin{aligned} f(x) &= P(R_i(X, Y) = x) \\ &= \sum_j \binom{N_i}{j}\binom{N_i}{j + N_i x}\binom{N - 2N_i}{2j + (x - 1)N_i} \bigg/ \binom{N}{N_i} \end{aligned}$$

for $x = -1, -1 + 1/N_i, -1 + 2/N_i, \ldots, +1$. For N even, $R_{[N/2]} = R_{[(N+1)/2]} = R_q$; for N odd, however, $R_{[N/2]}$, $R_{[(N+1)/2]}$, and R_q may differ slightly but $R_{[N/2]}$ and $R_{[(N+1)/2]}$ have the same distribution, which is asymptotically equivalent to the distribution of R_q.

In conclusion, we have defined a maximum deviation type nonparametric correlation coefficient R_g for use in testing the hypothesis of independence between two random variables. Moreover, R_g could be considered as a generalization of the quadrant correlation coefficient, R_q. Furthermore the power of R_g falls among that of other well-known nonparametric correlation coefficients when the sample comes from a bivariate normal, is as good as R_s for larger sample sizes of a bivariate exponential population, and is greater than that of the others when the population is a bivariate normal contaminated with biased outliers and the sample sizes are large. In addition, if a sample is severely biased in one of the tails (or, equivalently, the correlation is reversed from the bulk of the data in one of the tails), then R_g senses the correlation in the bulk of the data best. Thus R_g may be especially useful in problems in which outliers are present, contaminated populations are involved, or certain types of nonlinearity occur in bivariate data.

There are possible uses for R_g beyond just independence testing. For example, some forms of cluster analysis depend on the correlation coefficient as a measure of distance. This new coefficient used on nonnormal data could possibly cluster the data in a more attractive manner.

[*Received May 1985. Revised August 1986.*]

REFERENCES

Blomqvist, N. (1950), "On a Measure of Dependence Between Two Random Variables," *Annals of Mathematical Statistics*, 21, 593–600.

Blum, J. R., Kiefer, J., and Rosenblatt, M. (1961), "Distribution-Free Tests of Independence Based on the Sample Distribution Function," *Annals of Mathematical Statistics*, 32, 485–498.

Conover, W. J. (1980), *Practical Nonparametric Statistics*, New York: John Wiley.

Hájek, J., and Šidak, Z. (1967), *Theory of Rank Tests*, New York: Academic Press.

Hollister, R. A. (1984), "A Correlation Coefficient Based on Maximum Deviation," unpublished Ph.D. dissertation, University of Montana, Dept. of Mathematical Sciences.

Kendall, M. G. (1938), "A New Measure of Rank Correlation," *Biometrika*, 30, 81–93.

Kruskal, W. (1958), "Ordinal Measures of Association," *Journal of the American Statistical Association*, 53, 814–861.

Marshall, A. W., and Olkin, I. (1967), "A Multivariate Exponential Distribution," *Journal of the American Statistical Association*, 62, 30–44.

Rényi, A. (1959), "On Measures of Dependence," *Acta Mathematica Academiae Scientiarum Hungaricae*, 10, 441–451.

Schweizer, B., and Wolfe, E. F. (1981), "On Nonparametric Measures of Dependence for Random Variables," *The Annals of Statistics*, 9, 879–885.

Spearman, C. E. (1904), "The Proof and Measurement of Association Between Two Things," *American Journal of Psychiatry*, 15, 72–101.