Journal of Modern Applied Statistical Methods
November 2008, Vol. 7, No. 2, 416-434

1538 – 9472/08/$95.00

Constructing Confidence Intervals for Spearman's Rank Correlation with Ordinal Data: A Simulation Study Comparing Analytic and Bootstrap Methods

John Ruscio
The College of New Jersey

Research shows good probability coverage using analytic confidence intervals (CIs) for Spearman's rho with continuous data, but poorer coverage with ordinal data. A simulation study examining the latter case replicated prior results and revealed that coverage of bootstrap CIs was usually as good or better than coverage of analytic CIs.

Key words: Spearman's rank correlation, confidence intervals, bootstrap.

Introduction

Spearman's (1904) rank correlation[1] (r_S) is a nonparametric statistic that allows an investigator to describe the strength of an association between two variables X and Y without making the more restrictive assumptions of the Pearson product-moment correlation (r). To calculate r_S, one converts each variable to ranks, assigning equal ranks to any tied scores (but see Gonzales & Nelson, 1996, for alternative approaches to handling ties), and then uses the usual formula for r or this computational shortcut

$$r_S = 1 - \frac{6\sum d_i^2}{N(N^2-1)}, \qquad (1)$$

where the d_i are the differences in the ranked scores on X and Y for each pair of cases and N is the sample size. Because this statistic is sensitive only to the order of differences between adjacent scores, and not their magnitudes, it belongs to the family of ordinal statistics (Cliff, 1996).

Cliff (1996) argues that ordinal statistics such as r_S are better able to answer ordinal research questions than more conventional parametric statistics. For example, asking whether higher self-esteem is associated with higher academic achievement poses an ordinal question. Using r to address it requires assumptions that may be unrelated to the research question and can be difficult to satisfy. Whereas r measures the strength of a linear relationship between X and Y, r_S assesses how well an arbitrary monotonic function describes the relationship. Testing for the strictly linear relationship between self-esteem and academic achievement will underestimate the strength of a relationship if it is nonlinear. Also, the insensitivity of r_S to monotonic transformations of the data can be a significant strength when it is safer to presume a monotonic relationship between one's measure of a variable and the underlying construct than to presume a linear relationship (Cliff, 1996). Whereas r assumes bivariate normality, r_S makes no assumptions about the distribution of either variable. Wilcox (2003) discusses the sensitivity of parametric statistics to extreme scores and, in many instances, even small departures from their assumptions. Caruso and Cliff (1997) suggest that r_S should be less sensitive to extreme scores and a more inferentially robust measure than r.

In addition to the fact that r_S does not require assumptions of linearity or bivariate normality, r_S can be used with ordinal data. According to Stevens (1946), a variable is classified as ordinal if scores can be scaled as rank-ordered categories but the absolute distances between them are unknown. Cliff (1996) observed that many variables of interest to psychologists are ordinal in nature. When one

John Ruscio is an Associate Professor in the Department of Psychology. Email him at ruscio @tcnj.edu.

or both of a pair of variables is ordinal, using r_S enables researchers to study relationships using variables that do not meet the interval scaling requirement of *r*.

Methods for evaluating the statistical significance of r_S are based on its sampling distribution under the null hypothesis (H_0) of $\rho_S = 0$. A randomization test (Edgington, 1987) may be the best way to test H_0, and many textbooks present tables of critical values for relatively small sample sizes (e.g., critical values in Zar, 1972, have been reprinted). With sufficiently large samples, one can use an approximation to the *t* distribution with *df* =*N*–2:

$$t = \frac{r_S}{\sqrt{(1-r_S^2)/(N-2)}}. \quad (2)$$

This is the same approximation that is ordinarily used to test the statistical significance of *r*.

Although null hypothesis significance testing remains popular in the social and behavioral sciences, guidelines provided by the APA's Task Force on Statistical Inference (Wilkinson et al., 1999) and its *Publication Manual* (American Psychological Association, 2009) recommended constructing a confidence interval (CI) instead. This is usually more informative because a CI allows an assessment of the null hypothesis (i.e., if the CI includes 0, one would retain H_0, otherwise one would reject H_0) and provides additional information, such as the precision with which a population parameter has been estimated. The more narrow the CI, the greater the precision of the estimate.

Testing the statistical significance of r_S is possible because the sampling distribution under H_0 is asymptotically normal and the variance of r_S can be estimated as $1/(N-1)$ (Higgins, 2004). To construct a CI, however, one cannot assume that $\rho_S = 0$, and when $\rho_S \neq 0$ the variance of r_S is more complex. Techniques have been developed to estimate the variance of Fisher-transformed r_S such that, when transformed back into r_S units, the coverage of CIs constructed in this manner will approximate the nominal level. Several approaches have been developed and studied, and each is an adjustment to the technique used with *r*. After Fisher-transforming *r*, where $z_r = \tanh^{-1}(r)$, the usual estimate of the variance of z_r is $1/(N-3)$. With this estimate of the sampling error of z_r and the assumption that these errors are normally distributed, one can construct a CI as follows:

$$CI(\rho) = \tanh\left[z_r \pm \sqrt{\frac{1}{N-3}}(z_{(1+CL)/2}) \right], \quad (3)$$

where *CL* is the desired confidence level (e.g., .95) and $z_{(1+CL)/2}$ is the percentile point of a standard normal distribution below which the subscripted proportion of scores lies. For example, constructing a 95% CI for $r = .50$ and $N = 50$ would proceed as follows: $z_r = \tanh^{-1}(.50) = .5493$, $z_{(1+CL)/2} = z_{.025} = 1.96$, and $CI(\rho) = \tanh(.5493 \pm .1459 \times 1.96) = .26$ to .68. Note that for $r \neq 0$, this technique yields a CI asymmetric about *r*.

To construct a CI for ρ_S in a parallel fashion, one begins with the Fisher transformation $z_{r_S} = \tanh^{-1}(r_S)$ and then uses its estimated variance in much the same way shown in Eq. 3. Whereas the *z* distribution is used to form CIs for ρ, Woods (2007) recommended using the *t* distribution (with $df = N-2$) to form CIs for ρ_S. Because Woods found that the observed coverage of CIs for ρ_S often was below the nominal level, and sometimes substantially so, the *t* distribution will be used in the present study. (Using the *z* distribution would produce narrower CIs than using the *t* distribution, hence coverage even further below the nominal level.) Thus, the CI for ρ_S is constructed as follows:

$$CI(\rho_S) = \tanh[z_{r_S} \pm \sigma(z_{r_S}) \times t_{(1+CL)/2}], \quad (4)$$

with formulas to estimate $\sigma^2(z_{r_S})$, the variance of the Fisher-transformed r_S, developed by three sets of investigators: Fieller, Hartley, and Pearson (1957), Caruso and Cliff (1997), and Bonnett and Wright (2000). Each represents an ad hoc adjustment to the formula used to estimate the variance of z_r (recall that this is 1 / [N – 3]) that performed well under the conditions studied by its creators:

$$\sigma^2_F(z_{r_S}) = \frac{1.06}{N-3}, \quad (5)$$

$$\sigma^2_{CC}(z_{r_S}) = \frac{1}{N-2} + \frac{|z_{r_S}|}{6N+4\sqrt{N}}, \quad (6)$$

$$\sigma^2_{BW}(z_{r_S}) = \frac{1+r_S^2/2}{N-3}. \quad (7)$$

Caruso and Cliff (1997) studied CIs with ρ_S ranging from .00 to .89 using bivariate normal data with N = 10 to 200. Their technique (based on Eq. 6) achieved the nominal coverage levels. Bonnett and Wright (2000) studied CIs constructed using each of the three formulas shown above (Eqs. 5-7) with ρ_S ranging from .10 to .95 using bivariate normal data with N = 25 to 200. Their technique (Eq. 7) achieved good coverage even at large ρ_S (.80 to .95), where the other methods became liberal (i.e., coverage dropped below the nominal level). These results suggest that 95% CIs for ρ_S provide fairly accurate coverage for bivariate normal variables, with tendencies toward liberal coverage at large ρ_S and small N, and that the Bonnett and Wright formula for $\sigma^2(z_{r_S})$ may be the most useful of the three evaluated in these studies.

To date, only Woods (2007) investigated the coverage of CIs for ρ_S using ordinal data. Woods examined CIs constructed using each of the three formulas for $\sigma^2(z_{r_S})$ shown above using populations based on empirical data in which variables with either 4 or 5 categories correlated with one another from near-zero to large values of ρ_S; sample sizes in the simulation study ranged from N = 25 to 100. In the corrected results[2], Woods found that the Bonnett and Wright (2000) formula provided CIs with slightly better coverage than its rivals, but there remained room for improvement. For example, the coverage of nominally 95% CIs was below 90% for many conditions. Coverage dropped further below the nominal level for larger values of ρ_S, which is consistent with the findings of research using ratio scale data.

At least two factors that may constrain the performance of the analytic method of constructing a CI by using a formula for $\sigma^2(z_{r_S})$, at least under conditions that diverge from bivariate normality. First, each of the three formulas was developed as an ad hoc adjustment to the formula for estimating the variance of z_r. Because data may diverge substantially from bivariate normality (e.g., ordinal data will not be distributed in this way), it may not be possible to adjust the formula for the variance of z_r in a way that works well for a broad variety of data conditions. Second, constructing CIs for ρ_S using any of these formulas involves an assumption about the shape of the sampling distribution that may not be satisfied. Specifically, the t distribution is used to construct the CI. Whenever the sampling distribution does not follow the t distribution, the coverage of these CIs may deviate from the nominal level.

Bootstrap methods for constructing CIs avoid both of these potential problems (Efron & Tibshirani, 1993). Rather than using a formula to estimate the variance of a statistic and making an assumption about the shape of its sampling distribution, one treats the available data as the best estimate of the population, draws random samples from it a large number of times (this is known as resampling, which provides what are called bootstrap samples), and calculates the statistic in each of these bootstrap samples. The distribution of the statistic across the bootstrap samples constitutes an empirical sampling distribution.[3] The empirical sampling distribution is generated without recourse to assumptions such as bivariate normality, no formula is needed to estimate the variance of the statistic in the relevant population, and no assumptions are made about the shape of the sampling distribution. The strengths - and weaknesses - of bootstrap methods involve their heavy reliance on the empirical data rather than standard parametric assumptions (Kline, 2005).

Once one has generated an empirical sampling distribution, CIs can be obtained in several ways. The simplest, although not always the best, method for constructing a bootstrap CI is to record the values of the statistic in the sampling distribution that span the desired proportion of results, with the remainder lying beyond the CI in equal proportions in both tails. For example, suppose a sample of N = 50 cases of ordinal data yielded r_S = .72. Treating these

data as the population of pairwise scores, one can draw cases at random (with replacement) to obtain a new sample of $N = 50$, calculate r_S in this bootstrap sample, and repeat this procedure B times, where B is the number of bootstrap samples. When this was done $B = 2{,}000$ times and the results were rank-ordered, values of $r_S = .53$ and .86 spanned the middle 95% of the empirical sampling distribution. These constitute the lower and upper limits of a 95% CI for ρ_S using what is called the percentile bootstrap method (Efron & Tibshirani, 1993).

The percentile bootstrap operates by sorting the B values in the empirical sampling distribution and identifying the CI limits as the values indexed at the positions $B \times \alpha_L$ (for the lower limit) and $B \times \alpha_U$ (for the upper limit), where α_L and α_U are calculated as follows:

$$\alpha_L = (1 - CL)/2, \tag{8}$$

$$\alpha_U = (1 + CL)/2. \tag{9}$$

If either position is not a whole number, the next whole number toward the end of the range is used (e.g., if $B \times \alpha_L = 47.6$ and $B \times \alpha_U = 1943.1$, the values at positions 47 and 1944 would be used). For many statistics, percentile bootstrap CIs provide good coverage. When empirical sampling distributions are asymmetric, however, the bias-corrected and accelerated (BCA) bootstrap method often provides better coverage (Chan & Chan, 2004; Efron & Tibshirani, 1993). The BCA bootstrap method calculates α_L and α_U as follows:

$$\alpha_L = \Phi\left(z_0 + \frac{z_0 + z_{(1-CL)/2}}{1 - a(z_0 + z_{(1-CL)/2})} \right) \tag{10}$$

$$\alpha_U = \Phi\left(z_0 + \frac{z_0 + z_{(1+CL)/2}}{1 - a(z_0 + z_{(1+CL)/2})} \right), \tag{11}$$

where Φ is the standard normal cumulative distribution function and z_0 and a index median bias and skewness, respectively. Formulas for the latter two values appear below.

$$z_0 = \Phi^{-1}\left(\frac{\#(r_S^* < r_S)}{B} \right), \tag{12}$$

where r_S is the correlation in the replication sample, r_S^* is a correlation in a bootstrap sample, # is the count function (applied across all bootstrap samples), and Φ^{-1} is the inverse standard normal cumulative distribution function. The closer r_S is to the median of the empirical sampling distribution, the closer the proportion in parentheses will be to .5 and the closer z_0 will be to 0.

$$a = \frac{\sum (r_{S(\cdot)} - r_{S(i)})^3}{6\left(\sum (r_{S(\cdot)} - r_{S(i)})^2\right)^{3/2}}, \tag{13}$$

where $r_{S(i)}$ is a jackknife value of r_S calculated using all but the ith case and $r_{S(\cdot)}$ is the mean of all jackknife values. As is evident in the form of Eq. 13, a is related to skewness and indexes what is referred to in the bootstrap literature as acceleration, or the rate of change in the standard error of a statistic relative to its true parameter value. When $a = z_0 = 0$, Eqs. 10 and 11 simplify to Eqs. 8 and 9, in which case the BCA bootstrap method yields the same CI as the percentile bootstrap method. When $a \neq 0$ or $z_0 \neq 0$, Eqs. 10 and 11 involve adjustments to the values of α_L and α_U.

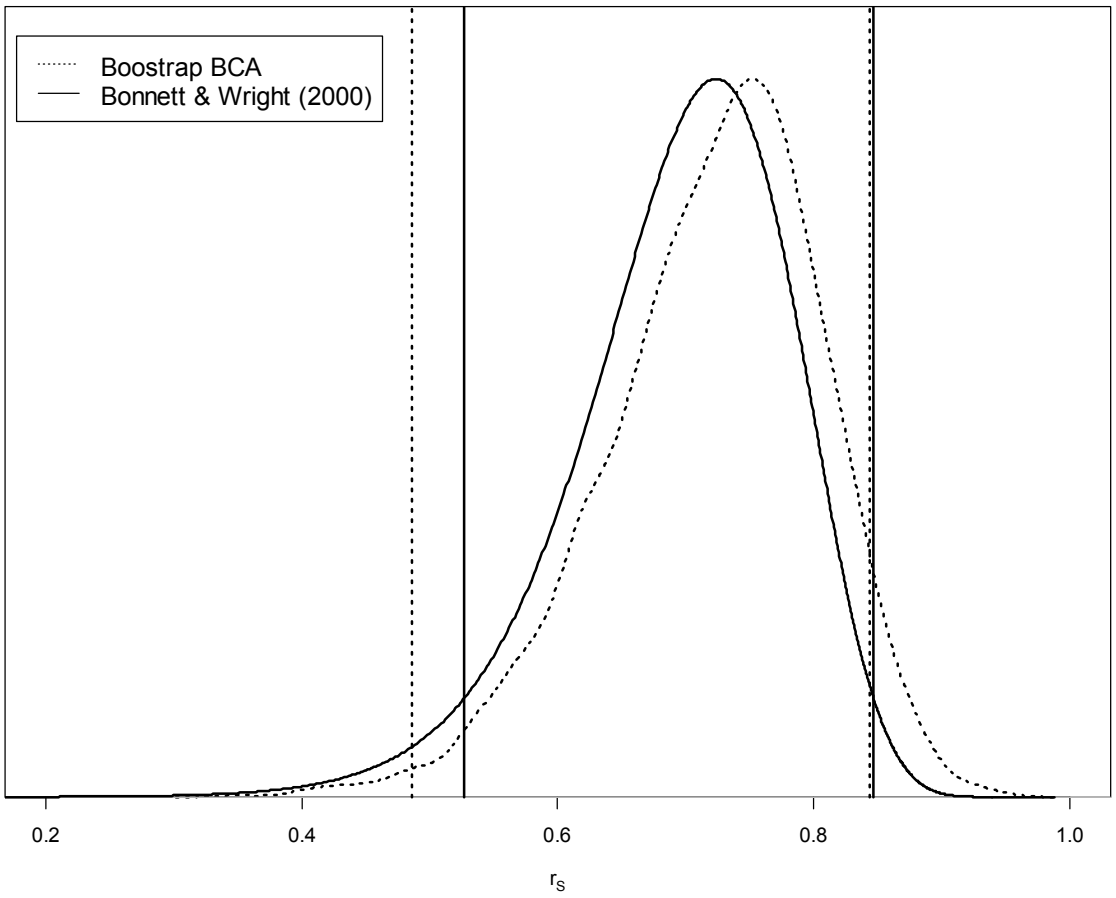
By indexing median bias and skewness to adjust α_L and α_U, BCA bootstrap CIs often provide better coverage than percentile bootstrap CIs. For example, in a study of CIs for ρ under conditions of range restriction, Chan and Chan (2004) found that the BCA bootstrap method yielded CIs with better coverage than did other bootstrap methods. Because the sampling distribution of Spearman's rank correlation is expected to be asymmetric when $\rho_S \neq 0$, the BCA bootstrap was included in the present study and the percentile bootstrap was not.

To illustrate the difference between conventional and bootstrap approaches, Figure 1 displays sampling distributions generated analytically, using the Bonnett and Wright (2000) estimate of $\sigma^2(z_{r_S})$, and empirically, using the BCA bootstrap method. Whereas the shape of the former is assumed (prior to

transformation from Fisher-transformed r_S back to ordinary r_S units, it followed the t distribution with 48 df), the latter is based on the observed results for B = 2,000 bootstrap samples drawn from the original data. The Bonnett and Wright 95% CI ranged from .53 to .85, which is nearly the same as the percentile bootstrap CI of .53 to .86. The BCA bootstrap method adjusted these limits downward, and this CI ranged from .49 to .84. Only the BCA bootstrap CI included the correct value of ρ_S = .50, so it appears that the adjustments for median bias (z_0 = -.085) and skewness (a = -.038) were helpful in this instance.

Because the construction of bootstrap CIs does not require a formula to estimate the standard error of r_S (or Fisher-transformed r_S) and does not assume the shape of the sampling distribution, it may provide better coverage than the analytic method for constructing CIs. On the other hand, bootstrap methodology for constructing CIs treats the sample data as the best estimate of the population and resamples from this bivariate distribution. Any irregularities in the sample can be magnified in bootstrap applications, and this can be especially problematic with small samples (Kline, 2005). The present study was designed to compare the coverage of analytic and bootstrap CIs for ρ_S across a wide range of ordinal data conditions, including small sample sizes.

Figure 1: Sampling distributions for r_S in analyses of a sample of N = 50 cases drawn from a population in which both variables were distributed asymmetrically across 5 categories and ρ_S = .50. Vertical lines represent the limits of 95% confidence intervals constructed from each sampling distribution.



Methodology

Design

Four factors were studied. First, the marginal frequencies of variables in the populations were either derived from empirical data or simulated using asymmetric, symmetric, or uniform distributions. Second, the size of the contingency table for a bivariate relationship was either 5 × 5 or 4 × 5, which limited each variable to a small number of ordered categories and allowed for equal or unequal numbers of categories. Third, ρ_S varied from zero to a very large value (.90). Fourth, sample size varied from small (N = 25) to modestly large values (N = 200).

Population Data

Four types of bivariate population distributions were included in the study. First, the distributions in Woods (2007) were used so that results for BCA bootstrap CIs could be compared to those for the methods in prior research. Because Woods focused primarily on measures of ordinal association in the gamma family, populations were selected such that Γ ranged from near zero (-.01 to .01) through small (.35 to .39), medium (.55 to .59), and large (.85 to .89) levels. Populations were not selected for values of ρ_S, and consequently these do not vary as widely or discretely as the four levels of Γ. At each level of Γ, the number of categories was selected such that variables had equal or unequal numbers of categories.

Specifically, both 5 × 5 and 4 × 5 contingency tables were used. Woods studied four sample sizes (N = 25, 50, 75, and 100), and each sample size had a corresponding population distribution from which cases were sampled (with replacement). The variables' marginal distributions generally were asymmetric. Figures 2 and 3 show the population distributions for all 32 conditions (4 sample sizes × 4 levels of Γ × 2 table sizes) in Woods' study.[4] In addition to Γ for each condition, ρ_S is shown. All samples drawn from the Woods populations had the same sizes as in the original study (N = 25, 50, 75, and 100).

Because Woods (2007) selected populations for study from an empirical data set, there is a degree of realism to the data conditions. However, the finite number of variable pairs available in these data may have precluded an orthogonal manipulation of the design factors. For example, marginal distributions are not independent of sample size or ρ_S. To supplement the distributions analyzed by Woods, three additional types of population distributions were created in which design factors were manipulated orthogonally. First, marginal distributions were similar to those used by Woods in that they were asymmetric.

Values for variables with 5 categories were sampled with probabilities of .55, .20, .12, .08, and .05; values for variables with 4 categories were sampled with probabilities of .60, .22, .11, and .07. These distributions approximated the asymmetry observed in many of Woods' populations. Second, marginal distributions were symmetric (and unimodal), with probabilities calculated using thresholds of -1.5, -.5, .5, and 1.5 in a standard normal distribution to create 5 categories and thresholds of -1, 0, and 1 to create 4 categories; these correspond to probability distributions of .07, .24, .38, .24, and .07 for 5 categories and .16, .34, .34, and .16 for 4 categories. Third, marginal distributions were uniform.

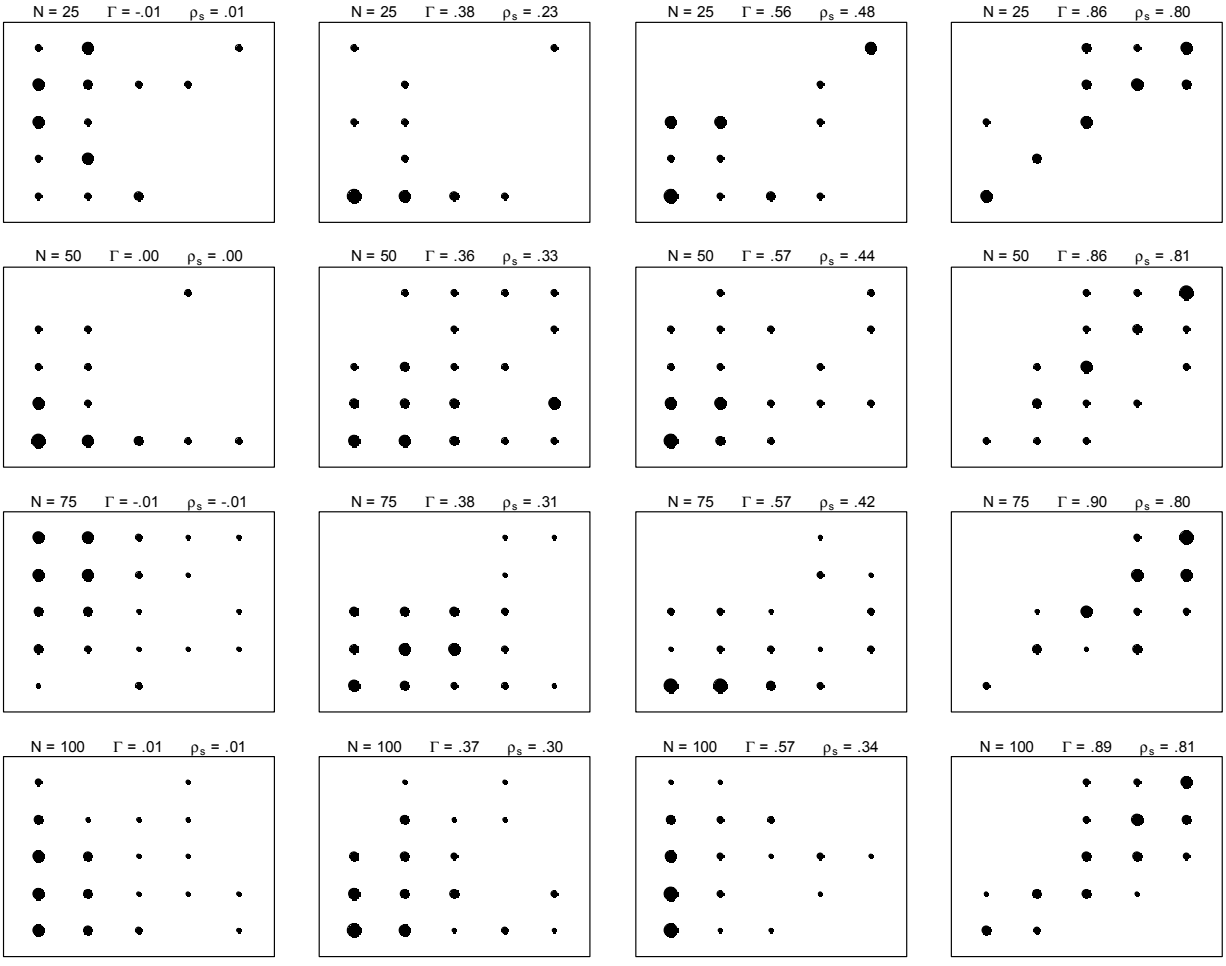
For each type of distribution, both 5 × 5 and 4 × 5 tables were created at each of six levels of ρ_S (.00, .10, .30, .50, .70, and .90). To generate each of these 36 bivariate population distributions (3 types of marginal distribution × 2 table sizes × 6 levels of ρ_S), the iterative technique developed by Ruscio, Ruscio, and Meron (2007), and subsequently generalized with improved efficiency by Ruscio and Kaczetow (2008), was used. This technique generates multivariate data sets with user-specified marginal distributions and correlation matrix. Both of the papers cited above demonstrate that this technique reproduces the desired distributions and correlations with good precision, especially at large sample sizes. In the present study, data were generated such that each of the 36 populations possessed 100,000 cases, which enabled a very close match between ρ_S as specified in the study design and as calculated in the finite population from which replication samples were drawn: With one exception, these values were within .005 of each other.[5] From each population, samples were

drawn with N = 25, 50, 100, and 200, yielding a total of 144 cells in this portion of the study design (36 populations × 4 sample sizes).

Replication Sample

Within each cell of the design, including the 32 conditions created by Woods (2007) and the 144 new conditions involving asymmetric, symmetric, and uniform populations, 1,000 replication samples were drawn for analysis. Whereas previous studies of CIs for ρ_S have used larger number of replication samples, this was not feasible in the present study due to the inclusion of a bootstrap method that required extensive resampling and analysis for each replication sample. For each replication sample, B = 2,000 bootstrap samples were drawn and analyzed. Using 1,000 replication samples per condition – the same number used in Chan and Chan's (2004) study of bootstrap CIs for ρ in situations of range restriction – was both feasible, given the inclusion of a computationally intensive bootstrap method, and adequate for informative comparisons among the four types of CI studied. Each replication sample was checked to ensure that the variance for each variable was greater than zero so that a correlation could be calculated. In a small number of instances, primarily when drawing small samples from asymmetric populations, all values for a variable were identical and that sample was not included in the study.

Figure 2: Population distributions for data conditions with 5 × 5 tables in Woods (2007). The area of each plotting symbol is proportional to the frequency in that cell of the contingency table.



Data Analysis

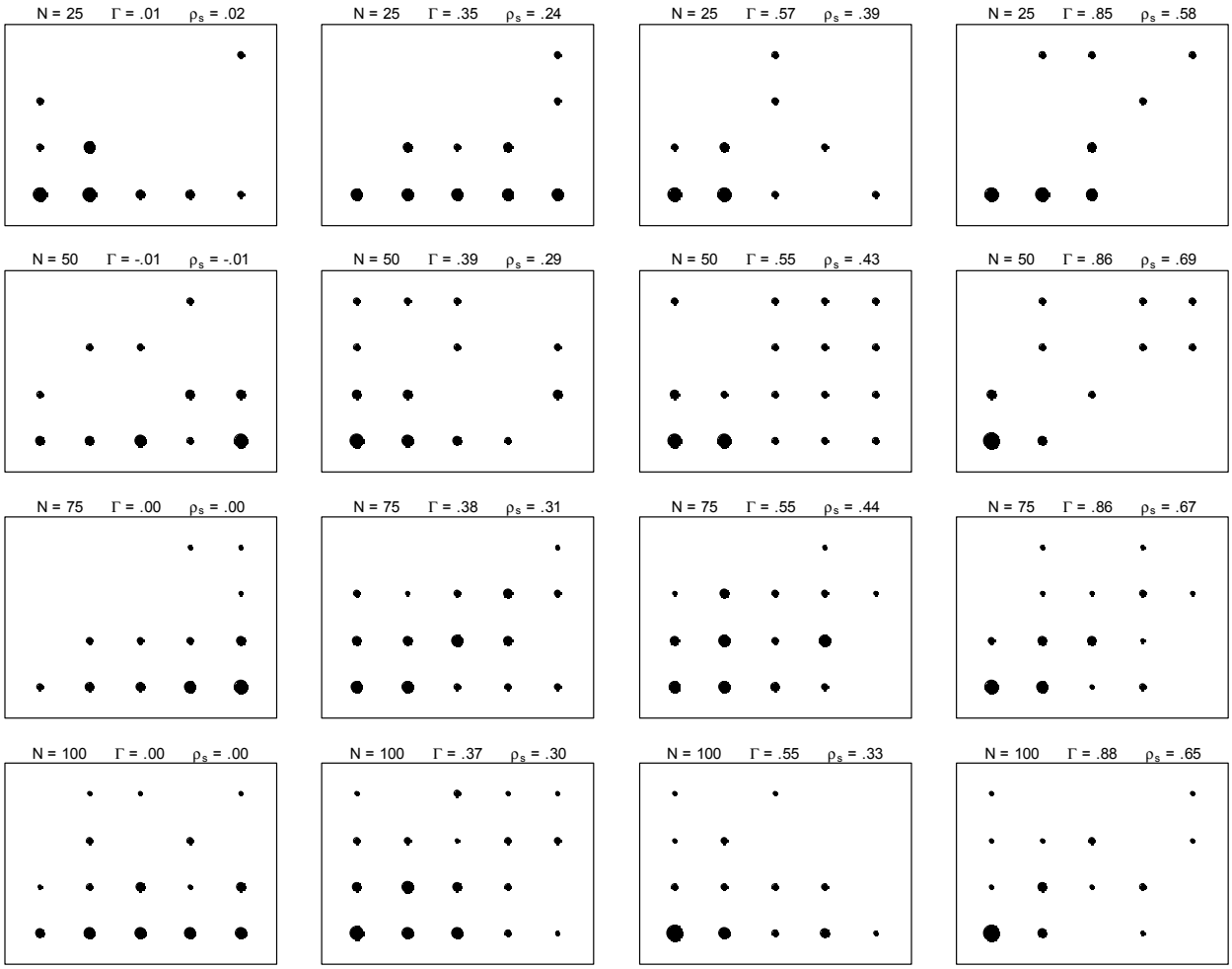
For each replication sample, r_S was calculated and Eqs. 5-7 were used to estimate the variance of $\sigma^2(z_{r_S})$ and construct CIs according to the methods of Fieller et al. (1957), Caruso and Cliff (1997), and Bonnett and Wright (2000). Then, B = 2,000 bootstrap samples - a quantity recommended by DiCiccio and Efron (1996) and also used by Chan and Chan (2004) - were drawn from each replication sample and r_S was calculated for each to construct a bootstrap BCA CI. The nominal level of all CIs was .95 (95%). Each bootstrap sample was checked to ensure that a correlation could be calculated (i.e., that both variables' variances were greater than zero); in a small number of instances, a new sample was drawn to replace one that was discarded because a correlation could not be calculated.

Within each cell of the design, observed coverage was recorded as the proportion of the CIs that included ρ_S (the value observed in the finite population from which replication samples were drawn). The absolute deviance between nominal and observed coverage was also recorded for each cell.

Results

Figure 4 displays the mean absolute deviance between nominal and observed coverage ($\overline{D}$) for each of the four types of CI. These graphs

Figure 3: Population distributions for data conditions with 4 × 5 tables in Woods (2007). The area of each plotting symbol is proportional to the frequency in that cell of the contingency table.



aggregate the results within types of population for all conditions, for each table size, for each level of ρ_S, and for each level of N. For the populations studied by Woods (2007), displayed in the upper-left panel, the results for the three types of analytic CIs are comparable to those in the original study; minor discrepancies are attributable to sampling variation between studies. $\overline{D}$ increased across levels of ρ_S for the analytic methods, reaching substantial values when ρ_S was large.

Because values of ρ_S did not vary discretely across the four levels in the design (recall that, strictly speaking, these were levels of Γ, not ρ_S), results were replotted in Figure 5 as observed coverage levels by ρ_S. This graph shows more clearly the tendency for coverage to fall below the nominal level with larger values of ρ_S. Relative to the coverage observed for the analytic methods, coverage for the bootstrap method was as good or better under most conditions, and much better for $\rho_S > .50$. Coverage for the bootstrap CIs remained within the control limits - the expected range of coverage results at $\alpha = .05$ with 1,000 replication samples, which is [.9365, .9635] - at even for the largest values of ρ_S. As expected, the bootstrap method yielded its largest values of $\overline{D}$ with the smallest samples ($N = 25$). Figure 5 shows that coverage for bootstrap CIs was outside of the control limits for only 4 of the 32 data conditions, each of which corresponded to an instance when $N = 25$. Different conditions seem to impair the performance of CIs for ρ_S constructed using analytic methods - in which case coverage falls below the nominal level as ρ_S increases - and the bootstrap method - in which case coverage is more erratic with smaller N.

Results for asymmetric populations (Figure 4) follow the same general pattern observed for the Woods (2007) populations. Here, the orthogonal manipulation of design factors helps to disentangle the effect of increasing ρ_S from the effects of different marginal distributions. As ρ_S increased, coverage remained closer to the nominal level for the bootstrap method than for the analytic methods; the difference was slight to nonexistent at $.00 \leq \rho_S \leq .30$, modest at $\rho_S = .50$, substantial at $\rho_S = .70$, and very large at $\rho_S = .90$. Once again, larger values of $\overline{D}$ were observed when the bootstrap method was used with smaller samples ($N = 25$) than with larger sample sizes ($50 \leq N \leq 200$).

With symmetric and uniform populations (Figure 4), perhaps the most striking result is that coverage for all methods approximated nominal levels fairly well under most conditions. Relative to the results for asymmetric populations, each of the methods achieved comparable or lower values of $\overline{D}$ under all conditions studied; note that that scaling of the y axes was held constant across panels in Figure 4 to facilitate this comparison. Nonetheless, the pattern of results across levels of ρ_S was similar to that observed for other populations: The bootstrap method maintained good coverage levels even at the highest values of ρ_S, whereas the analytic methods did not.

So far, results have focused primarily on absolute differences between observed and nominal coverage levels, and these discrepancies were averaged across cells in the design. To put more flesh on the bones of these results, for each CI method within each cell of the design, coverage was classified into one of seven categories using the control limits for $\alpha = .05$ (specified earlier), control limits of [.9322, .9678] for $\alpha = .01$, and control limits of [.9273, .9727] for $\alpha = .001$.

This classification indicates whether coverage was within all control limits, liberal (observed coverage less than the nominal level) to one of three extents ($\alpha = .05$, $\alpha = .01$, or $\alpha = .001$), or conservative (observed coverage greater than the nominal level) to one of these three extents. Figure 6 displays the results for the Woods (2007) populations, with results for each CI method in each cell of the design symbolized as within control limits (solid circle), liberal (downward-pointing triangles), or conservative (upward-pointing triangles); the size of a triangle corresponds to the most extreme α level at which the results fell beyond the control limits, with larger triangles indicative of greater deviance between observed and nominal coverage levels. Table 1 summarizes these results by tallying the frequency with which results fell into each of the seven

Figure 4: Mean absolute deviation between nominal (.95) and observed coverage.

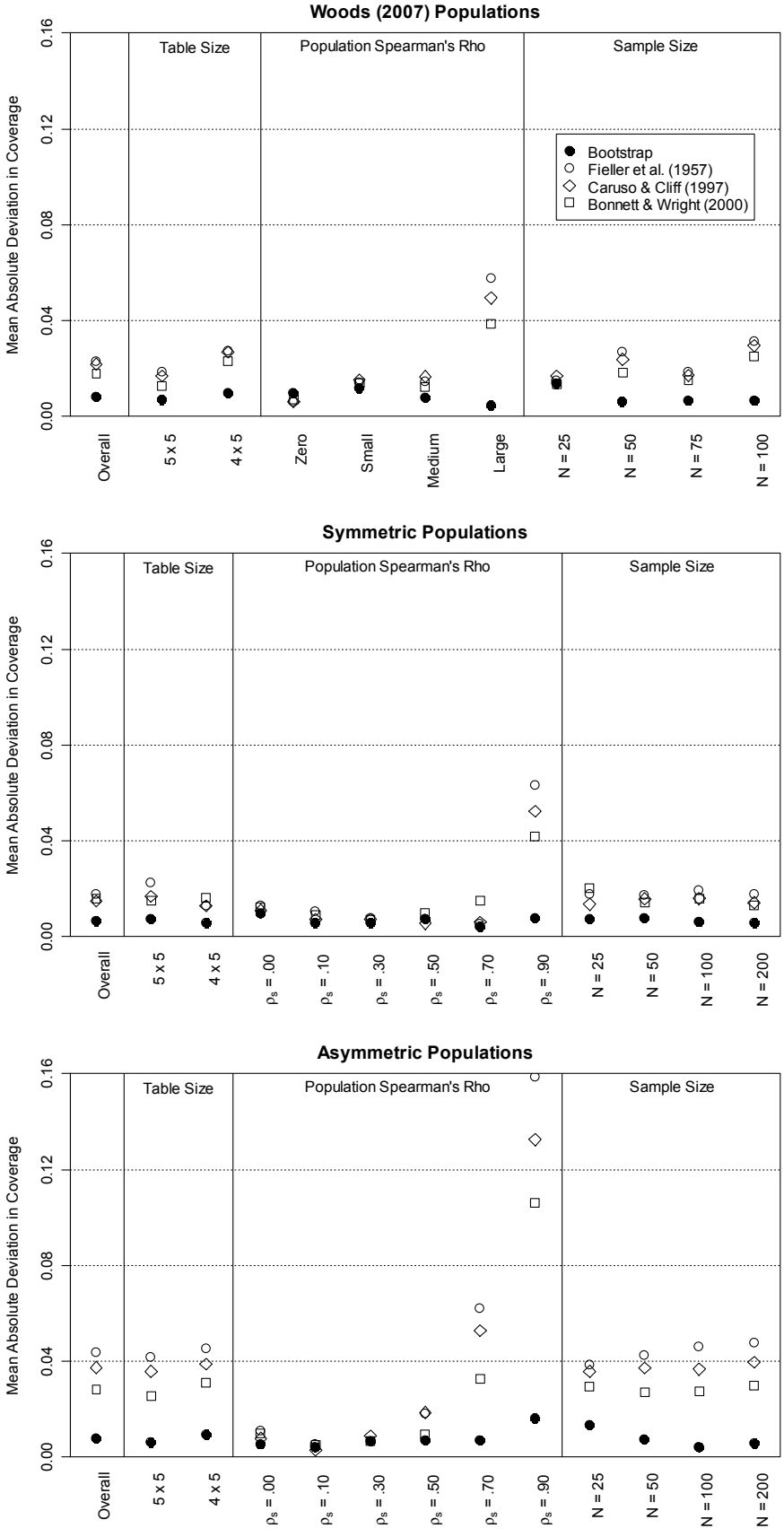
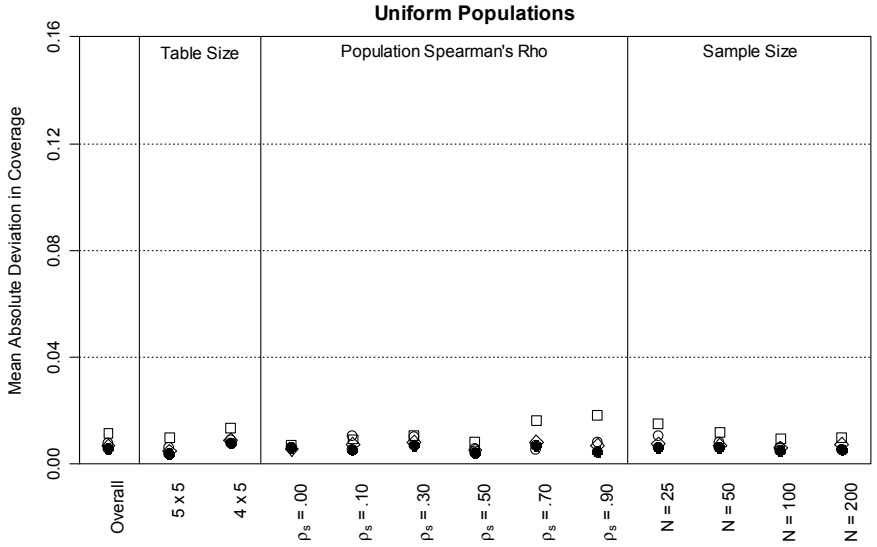


Figure 4 (continued): Mean absolute deviation between nominal (.95) and observed coverage.



categories for each CI method and population type.

Whereas the bootstrap method provided CIs whose coverage was within control limits for α = .05 88% of the time (28 of 32 conditions), the analytic methods provided CIs whose coverage was within these limits only 50% to 53% of the time. As noted earlier, the 4 exceptions for the bootstrap method occurred when N = 25 and exceptions for the analytic methods occurred more often as ρ_S increased. Figure 7 displays the results for the asymmetric, symmetric, and uniform populations, and Table 1 summarizes these results as tallied frequencies. The bootstrap method provided CIs whose coverage was within control limits for 92%, 85%, and 94% of the conditions in these three types of populations, respectively. The corresponding figures for the analytic methods were lower, often substantially lower, coverage erred on the liberal side two to three times as often as it erred on the conservative side, and most deviances exceeded even the α = .001 level. Across all populations and data conditions (i.e., all 176 cells of the study design), the bootstrap method provided CIs whose coverage was within control limits 90% of the time, whereas the figures for analytic methods were 64% (Fieller, et al., 1957), 67% (Caruso & Cliff, 1997), and 56% (Bonnett & Wright, 2000).

One potential explanation for the generally liberal coverage of the analytic methods is that r_S is a biased statistic, usually underestimating the value of ρ_S (Cliff, 1996). To the extent that r_S is a biased estimator of ρ_S, it should not be surprising that CIs constructed around this statistic do not contain the population value sufficiently often to attain the nominal coverage level. In the present study, however, the magnitude of bias was rather small. The mean level of bias ($r_S - \rho_S$) was calculated across the 1,000 replication samples within each of the 176 cells of the design, and the distribution of these values is shown in Figure 8 (M = -.0024, Mdn = -.0020). It seems unlikely that such a slight bias contributed substantially to the deviance between observed and nominal coverage levels for the analytically derived CIs. Instead, the two factors identified earlier - ad hoc formulas for estimating $\sigma^2(z_{r_S})$ and the use of the t distribution in constructing the CI - remain plausible candidates for the source of this deviance.

Conclusion

This article reveals some important similarities and differences in the coverage of CIs for ρ_S with ordinal data constructed using four methods

Table 1: Frequencies of Observed Coverage Levels Within and Beyond Control Limits.

CI Method	Population Type	− − −	− −	−	CL	+	+ +	+ + +
Bootstrap	Woods (2007)	2	0	1	28	1	0	0
	Asymmetric	2	0	0	44	1	1	0
	Symmetric	0	0	3	41	2	1	1
	Uniform	0	0	0	45	2	1	0
	All Populations	4	0	4	158	6	3	1
Fieller, et al. (1957)	Woods (2007)	9	3	0	16	1	1	2
	Asymmetric	18	2	4	21	2	0	1
	Symmetric	4	0	1	32	3	3	5
	Uniform	0	0	0	44	1	2	1
	All Populations	31	5	5	113	7	6	9
Caruso & Cliff (1997)	Woods (2007)	9	2	0	16	1	2	2
	Asymmetric	19	3	2	23	0	1	0
	Symmetric	4	0	0	35	2	2	5
	Uniform	0	0	0	44	2	2	0
	All Populations	32	5	2	118	5	7	7
Bonnett & Wright (2000)	Woods (2007)	6	0	4	17	1	2	2
	Asymmetric	14	2	3	27	1	0	1
	Symmetric	4	0	0	25	5	9	5
	Uniform	0	0	0	29	8	8	3
	All Populations	24	2	7	98	15	19	11

Notes: There were 32 data conditions for the Woods (2007) populations and 48 data conditions for each of the other three populations (asymmetric, symmetric, and uniform), for a total of 176 data conditions. − − − = coverage < .95 at $\alpha = .001$; − − = coverage < .95 at $\alpha = .01$; − = coverage < .95 at $\alpha = .05$; CL = coverage within control limits for .95 at $\alpha = .05$; + = coverage > .95 at $\alpha = .05$; + + = coverage > .95 at $\alpha = .01$; + + + = coverage > .95 at $\alpha = .001$.

Figure 5: Scatterplot of observed coverage by ρ_S for the Woods (2007) populations. Dashed lines show the control limits for nominal coverage of .95 at $\alpha = .05$, which are [.9365, .9635].

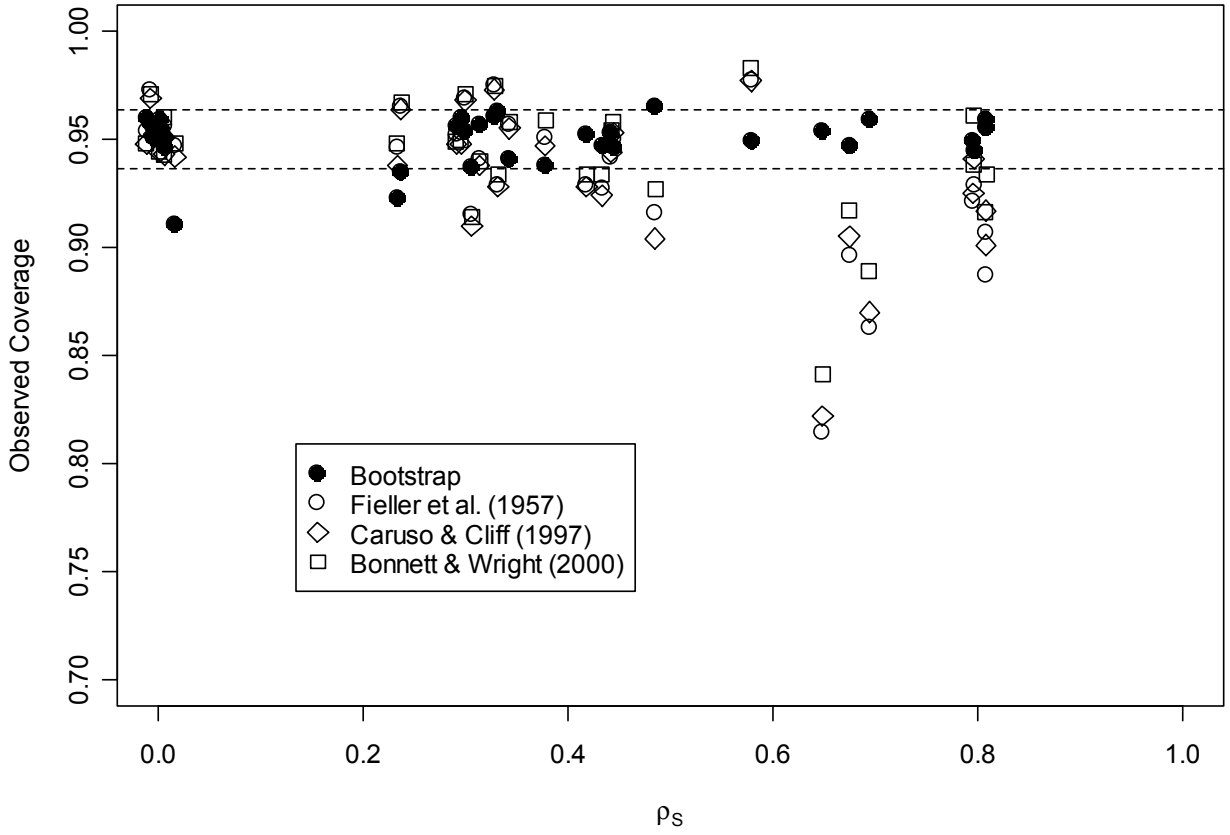


Figure 6: Chart indicates whether coverage was within the control limits of .95. These limits are [.9365, .9635] for α = .05, [.9322, .9678] for α = .01, and [.9273, .9727] for α = .001. B = bootstrap. F = Fieller, et al. (1957). CC = Caruso and Cliff (1997). BW = Bonnett and Wright (2000).

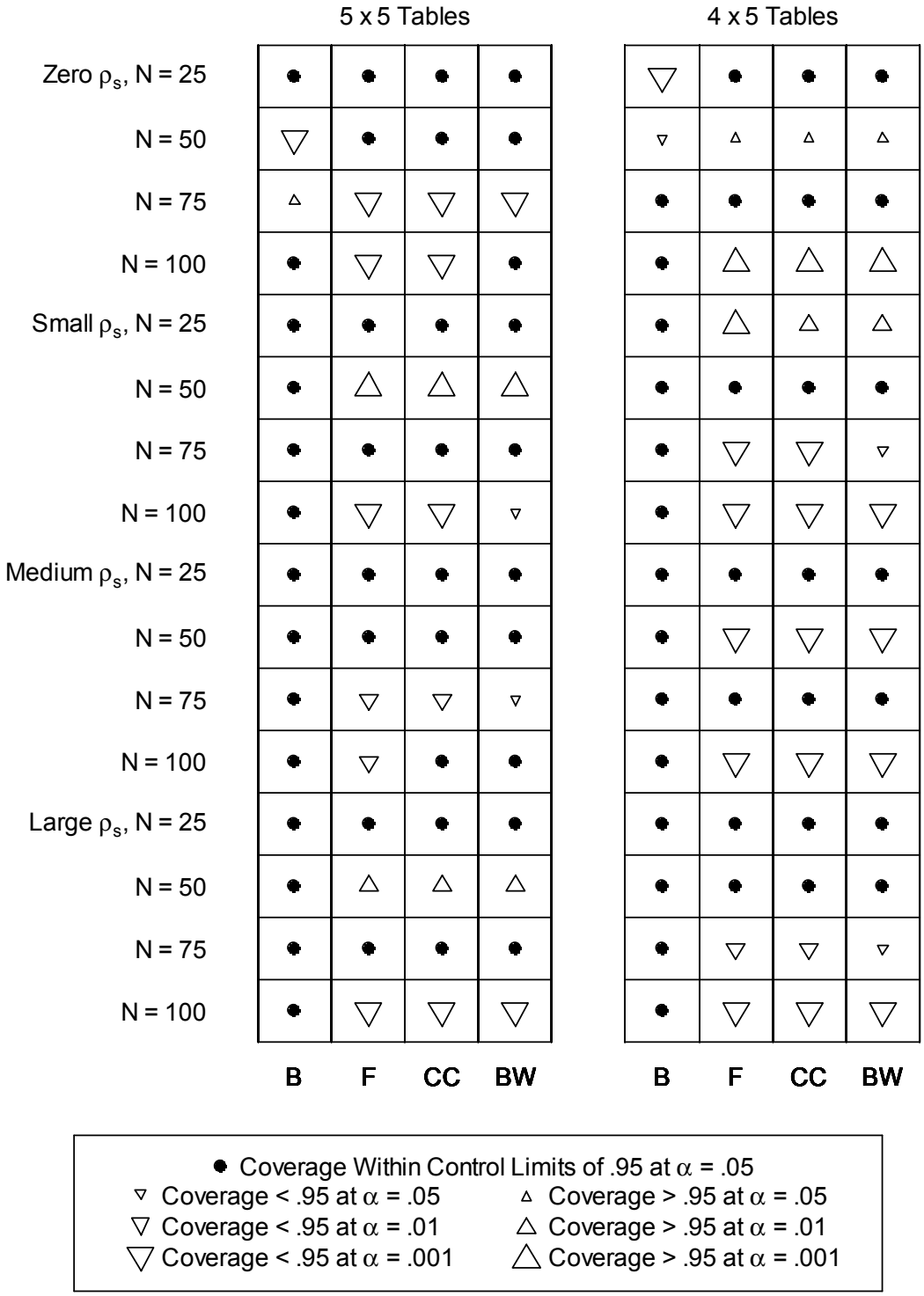


Figure 7: Chart indicates whether coverage was within the control limits of .95. These limits are [.9365, .9635] for α = .05, [.9322, .9678] for α = .01, and [.9273, .9727] for α = .001. B = bootstrap. F = Fieller, et al. (1957). CC = Caruso and Cliff (1997). BW = Bonnett and Wright (2000).

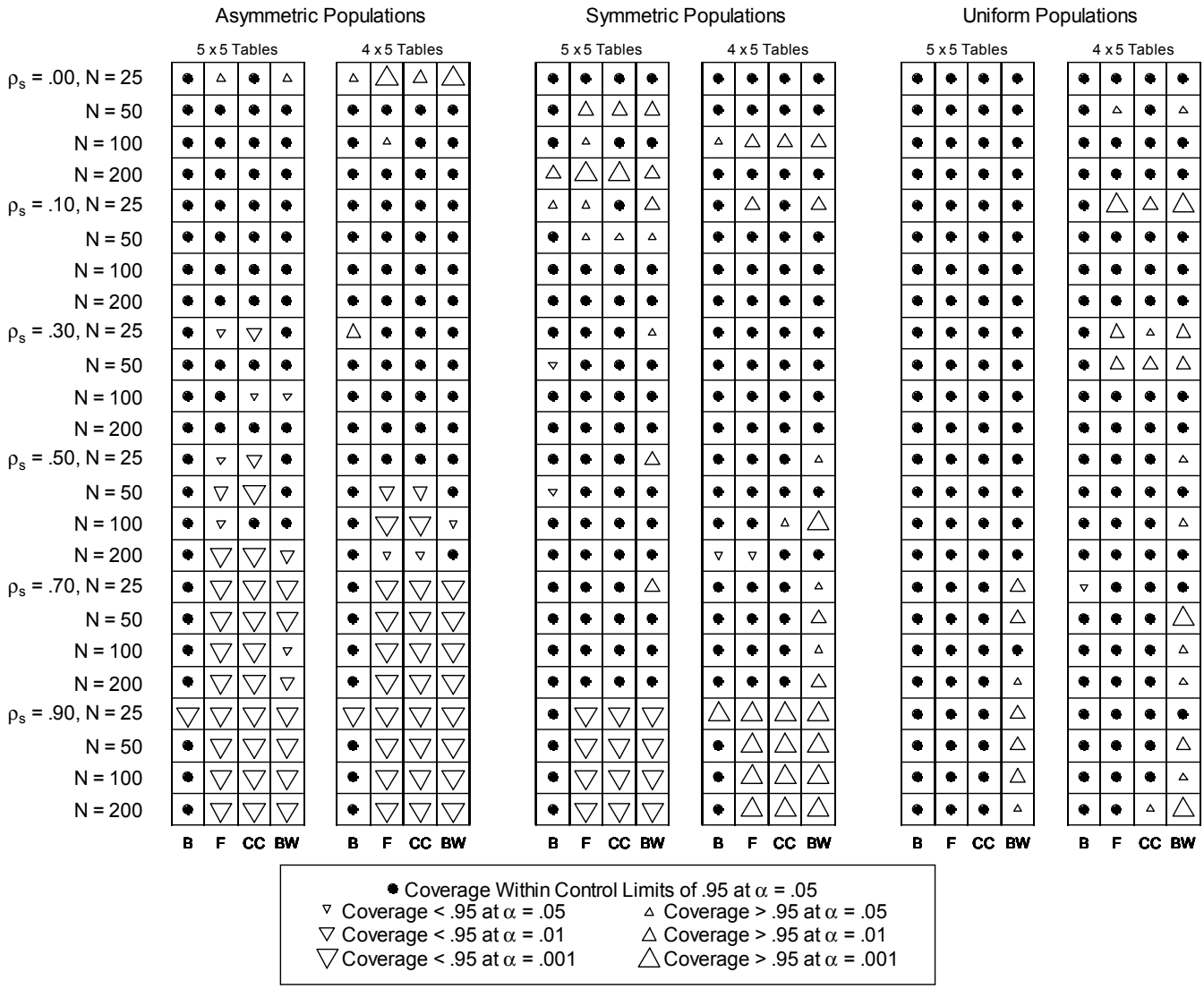
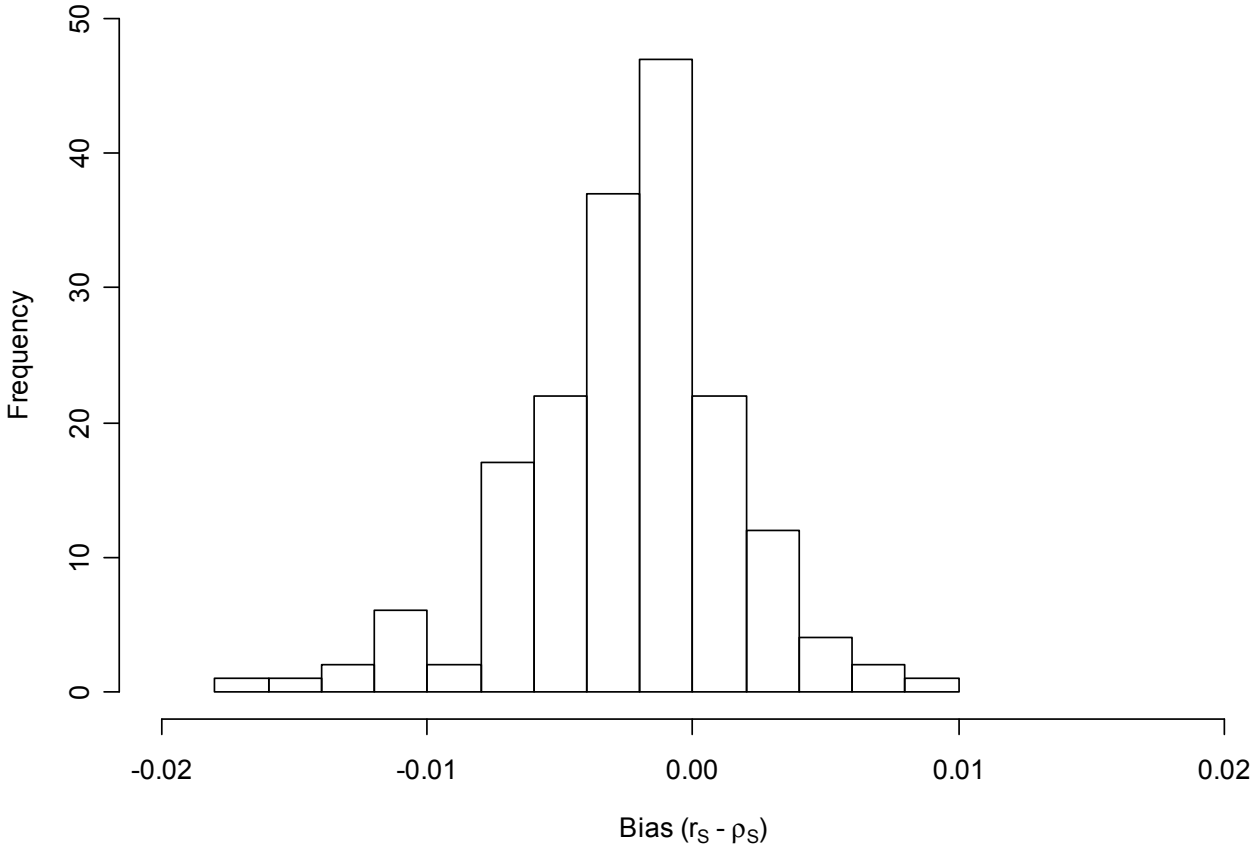


Figure 8: Histogram showing the bias in r_S as an estimator of ρ_S for all 176 cells of the study.



Under many conditions, both analytic and bootstrap methods provided CIs whose coverage approximated the nominal level of .95 well. These conditions included small values of ρ_S (between .00 and .30), moderate to large sample sizes (at least 50 cases), and symmetric (unimodal or uniform) marginal distributions. At larger values of ρ_S, the analytic methods tended to underestimate sampling error, yielding CIs that were too narrow and provided coverage less than the nominal level. This occurred for all marginal distributions studied, but the deviance was much greater for asymmetric than for symmetric distributions, and greater for unimodal than uniform distributions among those that were symmetric. Generally speaking, the BCA bootstrap method was robust across all values of ρ_S and each type of marginal distribution. To the extent that this method showed evidence of an Achilles' heel, it was the sometimes erratic coverage in the smallest samples studied ($N = 25$). Nonetheless, in many conditions with $N = 25$ and in nearly all conditions with $N \geq 50$, the BCA bootstrap method yielded CIs whose coverage was as good as or better than that of the analytic methods. At large values of ρ_S, this difference was substantial.

Although the study design spanned a broad array of data conditions - including several kinds of marginal distributions, sample sizes ranging from 25 to 200, and rank correlations ranging from .00 to .90 in ordinal data sets with relatively small numbers of categories - a number of issues remain to be clarified by future research. First, contingency tables of only two sizes were studied. Using Woods' (2007) investigation as a launching pad, the design included variables with either four or five categories crossed in 5 × 5 or 4 × 5 tables. With the exception of the symmetric, unimodal populations, 4 × 5 tables led to poorer coverage

than 5 × 5 tables. Because there are only two table sizes, it is impossible to determine whether this effect is due to the variables' unequal numbers of categories or due to the inclusion of a variable with fewer categories. Teasing apart these possibilities would require independently manipulating the number of categories for each variable and the equality vs. inequality of these numbers across variable pairs.

The use of only two table sizes also prohibits the generalization of results to either smaller or larger tables. At one extreme, it is possible to calculate r_S for two dichotomous variables. However, there are many other measures of association available for the analysis of 2 × 2 tables, each of which was developed to address a specific type of research question (for an overview, see Kraemer, Kazdin, Offord, Kessler, Jensen, & Kupfer, 1999). It seems unlikely that one would select r_S as the most appropriate measure for a 2 × 2 table, but there remain table sizes between 2 × 2 and 4 × 5 that merit further study.

Because the analytic methods studied here involve ad hoc adjustments to a technique developed for use with bivariate normal data, using them with increasingly small table sizes - which necessitate deviations from bivariate normality - is likely to lead to less satisfactory results. Bootstrap methods may be especially well-suited to these conditions, and this possibility should be studied. At the other extreme, ordinal data with increasingly large numbers of categories would approximate continuous distributions. As table sizes increase, it becomes possible for data to approximate bivariate normality more closely, and the difference in coverage between analytic and bootstrap CIs probably will depend on distributional forms. The present study suggests that the bootstrap holds important advantages with asymmetric distributions; whether or not this generalizes to larger table sizes should be studied.

Also worthy of investigation is the possibility that bootstrap methods might yield CIs for ρ_S with even better coverage if a larger number of bootstrap samples is used. In the present study, B = 2,000 bootstrap samples per replication sample were generated and analyzed both because this value is recommended in the bootstrap literature (e.g., DiCiccio & Efron, 1996; Efron & Tibshirani, 1993) and has been used in similar simulation studies (e.g., Chan & Chan, 2004) and because available computing resources made a value this large feasible in the context of the study design. Even though the BCA bootstrap method performed fairly well in an absolute sense, and as good as or better than the analytic alternatives under most conditions, there remains room for improvement. For example, across the 176 data conditions studied here, coverage for the bootstrap CIs was within the α = .05 control limits of the nominal coverage level only 90% of the time, not 95% of the time.

When using nonparametric bootstrap techniques such as the percentile or BCA methods, which locate the limits of CIs by indexing positions within an empirical sampling distribution, it is important to attain sufficient precision in the tails of this distribution. A larger value of B would help to flesh out these tails. Moreover, it should improve the estimates of the median bias (z_0) and acceleration (a) parameters that are used to adjust the positions for locating the lower and upper limits of the CI. Whereas z_0 may change relatively little with increasing B, a is akin to a skewness statistic and its sampling error is not trivial; larger values of B should be especially useful in obtaining better estimates of a. All of this takes on greater importance if one wishes to construct CIs with even higher confidence levels than the usual .95, which was used exclusively in this study. For example, using the percentile bootstrap method by locating the values that define the middle 99% of an empirical sampling distribution requires a very large value of B to stabilize its tails, which are defined by only .5% of bootstrap samples apiece (e.g., 10 samples in each tail for B = 2,000).

Even though there are fruitful areas for follow-up research and no method of constructing CIs for ρ_S can guarantee that the observed coverage will equal the nominal level under all data conditions, researchers who would like to use r_S to measure the association between two variables can be advised to calculate and report CIs. With at least a moderate sample size (e.g., $N \geq 50$), the bootstrap BCA method with B = 2,000 appears to provide good coverage levels

for any ρ_S from .00 to .90, even with as few as 4 or 5 ordered categories. If N is at least 25, the smallest value studied here, the analytic methods usually provided satisfactory coverage levels when ρ_S was not too large. For asymmetric distributions, coverage was good until ρ_S reached .50, and for symmetric distributions (unimodal or uniform), coverage was good until ρ_S reached .70. The only situations in which one would be well-advised to refrain from constructing CIs for ordinal data like those studied here are for small samples in which one's data are distributed asymmetrically and produce large values of r_S. Of course, conditions such as these would be extremely challenging for any correlational analysis - whether it involves testing H_0 or constructing a CI, using r_S or another measure of association - and it may be preferable to refrain from drawing strong conclusions from such data unless and until a method can be developed that handles them satisfactorily.

Notes

[1]The coefficient r_S is sometimes referred to as "Spearman's rho," which can be ambiguous in that Greek letters often are reserved for values calculated in populations rather than samples. In the present article, r_S will be used to denote the sample estimate of ρ_S, the population value of Spearman's rank correlation.

[2]Results for r_S published in Woods (2007) are superseded by those in a correction (Woods, 2008).

[3]Lee and Rodgers (1998) distinguished univariate and bivariate resampling for bootstrap applications with correlation coefficients. Whereas univariate resampling was found to be more useful for tests of statistical significance, it yields samples in which the marginal distributions reproduce those in the original data but the variables are uncorrelated (save for sampling error). As Lee and Rodgers note, bivariate resampling is required to construct CIs because this preserves not only the marginal distributions, but also the correlation in the original data. Thus, bivariate resampling was used exclusively for analyses presented in this paper.

[4]Categories were recoded to consecutive natural numbers. In the original populations used by Woods (2007), the coding of some variables began at 0 and others at 1, and some variables had frequencies of 0 at intermediate category numbers (e.g., scores of 0, 1, 2, and 4 occurred, with no scores of 3). Because this recoding preserved scores' rank order, it did not affect results.

[5]For the data condition with $\rho_S = .90$ and a 4×5 contingency table with symmetric marginal frequency distributions, ρ_S in the finite population of 100,000 cases was .8713. As in all other conditions, CI coverage was evaluated against the correlation observed in the finite population, not the correlation specified in the design, so the failure to generate a finite population with a .90 correlation should not bias the coverage results.

References

American Psychological Association. (2009). *Publication manual* (6th ed.). Washington, DC: Author.

Bonnett, D. G., & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall, and Spearman correlations. *Psychometrika*, *65,* 23-28.

Caruso, J. C., & Cliff, N. (1997). Empirical size, coverage, and power of confidence intervals for Spearman's rho. *Educational and Psychological Measurement*, *57,* 637-654.

Chan, W., & Chan, D. W. L. (2004). Bootstrap standard error and confidence intervals for the correlation corrected for range restriction: A simulation study. *Psychological Methods, 9,* 369-385.

Cliff, N. (1996). *Ordinal methods for behavioral data analysis.* Mahwah, NJ: Lawrence Erlbaum Associates.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science, 11,* 189-228.

Edgington, E. S. (1987). *Randomization tests.* New York: Marcel Dekker.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton, FL: CRC Press.

Fieller, E. C., Hartley, H. O., & Pearson, E. S. (1957). Tests for rank correlation coefficients: I. *Biometrika*, *44,* 470-481.

Gonzalez, R., & Nelson, T. O. (1996). Measuring ordinal association in situations that contain tied scores. *Psychological Bulletin*, *119,* 159-165.

Higgins, J. J. (2004). *An introduction to modern nonparametric statistics*. Pacific Grove, CA: Brooks/Cole.

Kline, R. B. (2005). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Kraemer, H. C., Kazdin, A. E., Offord, D. R., Kessler, R. C., Jensen, P. S., & Kupfer, D. J. (1999). Measuring the potency of risk factors for clinical or policy significance. *Psychological Methods*, *4,* 257-271.

Lee, W.-C., & Rodgers, J. L. (1998). Bootstrapping correlation coefficients using univariate and bivariate sampling. *Psychological Methods*, *3,* 91-103.

Ruscio, J., & Kaczetow, W. (2008). Simulating multivariate nonnormal data using an iterative approach. *Multivariate Behavioral Research*, *43*, 355-381.

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

Ruscio, J., Ruscio, A. M., & Meron, M. (2007). Applying the bootstrap to taxometric analysis: Generating empirical sampling distributions to help interpret results. *Multivariate Behavioral Research*, *44,* 349-386.

Stevens, S. S. (1946, June 7). On the theory of scales of measurement. *Science*, *103,* 677-680.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques.* San Diego, CA: Academic Press.

Wilkinson, L., and the APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54,* 594-604.

Woods, C. M. (2007). Confidence intervals for gamma-family measures of ordinal association. *Psychological Methods*, *12,* 185-204.

Woods, C. M. (2008). Correction to Woods (2007). *Psychological Methods*, *13*, 72-73.

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