Comparing Dependent Correlations

RAND R. WILCOX
TIAN TIAN
University of Southern California

ABSTRACT. In a recent article in *The Journal of General Psychology*, J. B. Hittner, K. May, and N. C. Silver (2003) described their investigation of several methods for comparing dependent correlations and found that all can be unsatisfactory, in terms of Type I errors, even with a sample size of 300. More precisely, when researchers test at the .05 level, the actual Type I error probability can exceed .10. The authors of this article extended J. B. Hittner et al.’s research by considering a variety of alternative methods. They found 3 that avoid inflating the Type I error rate above the nominal level. However, a Monte Carlo simulation demonstrated that when the underlying distribution of scores violated the assumption of normality, 2 of these methods had relatively low power and had actual Type I error rates well below the nominal level. The authors report comparisons with E. J. Williams’ (1959) method.

Keywords: dependent correlations, Monte Carlo simulation, power, Type I error

Consider three random variables—Y, X₁, and X₂—that have some unknown trivariate distribution. Let ρ₁ and ρ₂ be Pearson’s correlation between (a) Y and X₁ and (b) Y and X₂, respectively, and let ρ₁₂ be Pearson’s correlation between X₁ and X₂. Researchers in quantitative psychology have given considerable attention to the problem of testing \( H₀: ρ₁ = ρ₂ \).

Recently, Hittner, May, and Silver (2003) compared eight methods for accomplishing this goal, which had stemmed from Hotelling (1940), Olkin (1967), Dunn and Clark (1969), Meng, Rosenthal, and Rubin (1992), Steiger (1980; includes two modifications of Dunn and Clark’s method), Williams (1959), and Hendrickson, Stanley, and Hills’ (1970) suggestion of a modification of Williams’ method. Hittner et al. found that several methods performed reasonably well in terms of Type I errors when the marginal random variables have a normal or uniform distribution. When the marginal random variables have an exponential distribution, again all eight methods performed reasonably well.

*Software for applying the recommended method is available from Rand R. Wilcox.*

*Address correspondence to Rand R. Wilcox, Department of Psychology, University of Southern California, Los Angeles, CA 90089-1061, USA; rwilcox@usc.edu (e-mail).*
when $\rho_1 = \rho_2 = .10$ or .40. But when $\rho_1 = \rho_2 = .70$, all eight methods had estimated Type I error probabilities exceeding .10 when those researchers tested at the .05 level with a sample size of $N = 20$. The estimated probability of a Type I error was based on simulations with 2,000 replications.

Of particular concern is that when Hittner et al. (2003) increased the sample size to $N = 300$, the estimated probability of a Type I error again exceeded .10 for all eight methods with little or no improvement versus $n = 20$. Steiger (1980) documented concerns about Hotelling’s (1940) method. Because of Duncan and Layard’s (1973) reported theoretical results, the poor performance of the methods based in part on Fisher’s $z$ is not surprising. Duncan and Layard showed that Fisher’s $z$ is theoretically sound (i.e., uses a correct estimate of the standard error) under independence or with samples from normal distributions. But otherwise, as when sampling from skewed distributions or symmetric distributions with heavy tails (such as the contaminated normal), researchers are using an incorrect estimate of the standard error regardless of the sample size. Duncan and Layard illustrated that Fisher’s $z$ can be even more unsatisfactory than revealed by Hittner et al. when testing the hypothesis of a 0 correlation. For example, sampling from a contaminated normal, they found that the actual Type I error probability can exceed .30 when testing at the .05 level. Moreover, because an incorrect estimate of the standard is being used, the ability to control the probability of a Type I error does not improve as the sample size becomes large.

In the present study, our goal was to expand on the Hittner et al. (2003) research by considering a variety of new methods for offering better control over the probability of a Type I error. Although the importance of a Type I error can depend on the situation, Bradley (1978) has argued that in many instances, when researchers test at the .05 level, at a minimum the actual probability of a Type I error should be between .075 and .025. Based on Bradley’s criterion, we inferred that none of the methods examined by Hittner et al. is always satisfactory even with a sample size of 300. Among the new methods that we examined, three (Methods D1, D2, and B1) achieved the goal of a Type I error probability less than .075 among all the situations considered by Hittner et al. Method B1 also achieved the goal of having the actual Type I error probability greater than .025 among the situations considered by Hittner et al. However, its power can be poor relative to Methods D1 and D2, and we do not recommend it for general use.

Review of Williams’ (1959) $T$ Test

One of the two methods that Hittner et al. (2003) found to perform relatively well among the eight techniques that they studied was Williams’ (1959) $T$ test. The other was the Dunn and Clark (1969) method, which is based in part on Fisher’s $z$. Because of the theoretical and empirical results reported by Duncan and Layard (1973), we are confident in eliminating the Dunn and Clark method from consideration. To add perspective, we replicate the results for Williams’ test,
Wilcox & Tian 107

reported by Hittner et al. To make this article reasonably self-contained, we note that Williams’ test statistic is given by Equation 1, where \( n \) is the sample size, and \( r_1, r_2, \) and \( r_{12} \) are the usual estimates of \( \rho_1, \rho_2, \) and \( \rho_{12}, \) respectively, whereas \( r = (r_1 + r_2)/2 \) and \( |R| \) is the determinant of the correlation matrix, \( R. \)

\[
I = \frac{r_1 - r_2}{\sqrt{(n-1)(1+r_{12})} \left[ 2 \frac{n-1}{n-3} |R|^{-1} - (r_1^2 + r_2^2 - 2) \right]}^{1/2}
\]

Methods D1 and D2

In this section, we describe the two methods we found to perform best among all considered. To visualize Method D1, first imagine that the \( X \) variables have been standardized. In formal terms, let \( Z_1 = (X_1 - \mu_1)/\sigma_1 \) and \( Z_2 = (X_2 - \mu_2)/\sigma_2, \) where \( \mu_j \) and \( \sigma_j \) are the population mean and standard deviation associated with \( X_j (j = 1, 2), \) respectively. Let \( D = Z_1 - Z_2 \) and \( \rho_D \) be the correlation between \( D \) and \( Y. \) Then \( \rho_D = 0 \) if and only if the null hypothesis is true. To see this, note

\[
COV(Y, Z_1 - Z_2) = COV(Y, Z_1) - COV(Y, Z_2).
\]

Because \( Z_1 \) and \( Z_2 \) have a standard deviation of 1, dividing this last equation by \( \sigma \) yields the desired result. In practice, the population means and variances are not known, so that \( D \) is computed with the population means and variances replaced by their usual estimates. The strategy is simply to use the usual Student’s (e.g., Wilcox, 2003) \( T \) to test the null hypothesis, \( \rho_D = 0. \)

As we show, in terms of Type I errors, Method D1 performs well in most situations considered by Hittner et al. (2003), but there are instances in which it is too conservative. That is, the actual probability of a Type I error can drop well below the nominal level when both \( \rho_1 \) and \( \rho_2 \) are sufficiently close to 1. Method D2, which is based in part on Method D1, is designed to reduce this problem. The basic idea is to momentarily assume that the null hypothesis is true, determine an adjustment of the \( p \) value when sampling from a normal distribution, and then adjust the \( p \) value associated with Method D1 accordingly. But a fundamental problem is that an appropriate adjustment depends on the unknown values of \( \rho_{12}, \rho_1, \) and \( \rho_2. \)

The strategy is to estimate these parameters in the usual way, use simulations to determine an appropriate adjustment of the \( p \) value when the null hypothesis is true, and apply this adjustment to the \( p \) value associated with Method D1.

More precisely, generate \( n \) vectors of observations from a trivariate normal distribution having correlations \( r_1 = r_2 = r \) and \( r_{12}. \) Apply Method D1 to the observations that you just generated, and let \( p_1 \) be the \( p \) value. Now repeat this process \( B \) times and label the resulting \( p \) values as \( p_1, \ldots, p_B. \) Because the data were generated under a condition where the null hypothesis was true, when you test at the .05 level about 5\% of the \( p \) values should be less than or equal to .05. For illustrative purposes, imagine that 5\% of the \( p \) values are less than or equal to .10. This means that to achieve a Type I error probability of .05, you
should reject if the \( p \) value is less than or equal to .10. The idea is to determine via simulations the \( p \) value needed to achieve a .05 probability of a Type I error, and then adjust the \( p \) value of Method D1 accordingly. In particular, put the \( p_B \) values in ascending order, yielding \( p_1 \leq p_2 \leq \ldots \leq p_B \). Here, use \( B = 500 \), and estimate the \( p \) value that you should use so that the probability of a Type I error is .05 is \( p = p_{(25)} \). If \( p \) is the \( p \) value returned by Method D1, then an adjusted \( p \) value is given by \( p_{\text{adj}} = \frac{.05p}{p} \), in what we call Method D2.

**Methods Considered and Discarded**

We provide a brief outline of the other methods that we considered but found to be unsatisfactory. Because the methods are not recommended, we provide minimal details.

**Method B1.** The first alternative begins by using only the first half of the data to estimate \( \rho_1 \) and the other half to estimate \( \rho_2 \). The effect is to have independent estimates of the correlations, in which case you can use the method outlined by Wilcox (2003, p. 277), which has been found to control the probability of a Type I error fairly well under seemingly extreme departures from normality. The technique is based on a slight modification of the basic percentile bootstrap method. Briefly, for each group, pairs of observations are randomly resampled from the data with replacement, the difference between the correlations is noted, the process is repeated many times, and, with large sample sizes, the middle 95% of the estimated differences yields a confidence interval for \( \rho_1 - \rho_2 \). For small sample sizes, a simple adjustment is used. The concern is that not all of the data are being used to estimate \( \rho_1 \) and \( \rho_2 \), suggesting that power may be relatively low. This was found to be the case in the simulations that we will describe later in this article. In particular, power does not compare well with Methods D1 and D2.

**Method B2.** Method B2 was a basic percentile bootstrap technique designed specifically for dealing with regression (e.g., Wilcox, 2003, p. 495). The thought was that because this approach works well when comparing independent correlations, it deserved consideration for the problem at hand. Unfortunately, Type I error was unstable as a function of the distribution used to generate the data.

**Method B3.** An alternative to testing the hypothesis \( H_0: \rho_{dy} = 0 \) is to test \( H_0: \beta_1 = \beta_2 \), where \( \beta_1 \) and \( \beta_2 \) are the least squares regression slopes for (a) \( Z_1 \) and \( Y \) and (b) \( Z_2 \) and \( Y \), respectively. Researchers have found that certain variations of wild bootstrap methods perform well when they are dealing with least squares regression (e.g. Cribari-Neto, 2004; Flachaire, 2005; Godfrey, 2006). The method is based in part on generating (bootstrap) \( Y \) values (with the goal of estimating the null distribution) and on multiplying the residuals by a random variable that has
mean 0 and variance 1. Researchers have noted three variations for generating bootstrap values. The first is based in part on a standardized uniform random variable, and the other two use a discrete distribution (Cribai-Nato; Flachaire; Godfrey). We considered all three variations and found them to be unsatisfactory in terms of Type I errors.

Method B4. Another possibility is to proceed in the same manner as in Method D1, except by replacing Pearson’s correlation with Spearman’s rho or Kendall’s tau. But when Method D1 had Type I errors well below the nominal level, the same problem occurred when we used Spearman’s rho or Kendall’s tau.

Some Simulation Results

We used simulations to compare Methods D1 and D2 with Williams’ (1959) method. We generated data in the exact same manner used by Hittner et al. (2003). We let \( R \) be the correlation matrix and form the Cholesky decomposition \( U'U = R \), where \( U \) is the matrix of factor loadings of the principal components of the square-root method of factoring a correlation matrix, and \( U' \) is the transpose of \( U \). Next, we generated an \( n \times 3 \) matrix of data, \( X \), from a specified distribution (normal, uniform, or exponential). Then the matrix product \( XU \) produced an \( n \times 3 \) matrix of data that had the population correlation matrix \( R \). For each condition, we drew 2,000 samples of size \( N \) and used the proportion of rejections to estimate the actual Type I error probability when testing at the .05 level. We used sample sizes 20, 50, and 100. For \( N = 100 \), we did not obtain new insights, and so, for brevity, we do not report the results. We took the values of \( \rho_{12} \) as .1, .3, and .6 and the values of \( \rho_1 \) and \( \rho_2 \) as .10, .40, and .70.

For power, still following the approach used by Hittner et al. (2003), we took the magnitude of the difference between \( \rho_1 \) and \( \rho_2 \) (the effect sizes; ES) as .10, .30, and .60.

Results

Table 1 presents the estimated Type I error rates, where \( W \) indicates Williams’ (1959) method. Williams’ method performs well under normality, with the estimated Type I errors ranging between .040 and .059. It performs nearly as well for the uniform distribution. But for the exponential distribution, the estimates range between .038 and .107. The largest estimate among all three distributions, with Method D1, is .053. Thus, Method D1 avoids Type I errors well above the nominal level, but some of the estimates drop below .001. As for Method D2, the largest estimate is .059, and the smallest is .010. Method D2 performs better than Williams’ method in avoiding Type I errors well above the nominal level. Also, it improves on Method D1 in avoiding estimates well below the nominal level. In some cases, the improvement of Method D2 over Method D1 is striking.
For example, with \( N = 50 \) and \( \rho_1 = \rho_2 = .70 \), the estimates for Method D1—with corresponding \( \rho_{12} \) equal to .10, .30, and .60—are .002, .001, and .001, respectively. In contrast, for Method D2, the estimates are .033, .031, and .051.

Table 2 reports the estimated power. Williams’ (1959) test always has higher power than do Methods D1 and D2, and D2 always has higher power than does D1. Of course, this circumstance is not surprising when Williams’ test has a Type I error rate well above the nominal level. Even when Williams’ test and Method D2 have reasonably similar Type I error probabilities, Williams’ test is preferable. But the cost of using Williams’ test is that the actual Type I error may exceed twice the nominal level even with a fairly large sample size. In some situations, the improvement of D2 over D1 is large. For example—with \( N = 50 \), effect size of .30, and \( \rho_{12} = .10 \)—power for Methods D1 and D2 is approximately .30 and .052, respectively.

**Discussion**

Among the numerous methods for comparing dependent Pearson’s correlations, many methods perform poorly in certain situations, in terms of
Type I errors, and no method has been found that is completely satisfactory. However, we can make a recommendation depending on the extent to which psychologists want to avoid Type I error probabilities well above .05: Use Method D1 or D2. Method D1 is simple and avoids Type I errors well above the nominal level, but it can be too conservative and can have relatively low power. Method D2 is much more satisfactory, and writing computer software for applying it is easy. R and S-PLUS functions (twodcor10 and twodcor8) are available from Rand R. Wilcox on request. A possible criticism of Method D2 is that it uses simulations to determine the critical value. However, with the speed of modern computers, a large number of replications is possible so as to ensure an accurate estimate. Here, 500 replications were found to be satisfactory in terms of controlling Type I errors. If a researcher desires a much larger number of replications, it would be practical from the point of view of execution time.

There is a variety of new robust analogs of Pearson’s correlation beyond the classic Kendall’s tau and Spearman’s rho (e.g., Wilcox, 2003). Researchers know them to have certain advantages over Kendall’s tau and Spearman’s rho, and they

### TABLE 2. Estimated Power

<table>
<thead>
<tr>
<th>N</th>
<th>ES</th>
<th>( \rho_{12} )</th>
<th>Normal</th>
<th>Uniform</th>
<th>Exponential</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>W</td>
<td>D1</td>
<td>D2</td>
</tr>
<tr>
<td>20</td>
<td>.10</td>
<td>.10</td>
<td>.12</td>
<td>.01</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td>.10</td>
<td>.30</td>
<td>.12</td>
<td>.01</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>.10</td>
<td>.60</td>
<td>.15</td>
<td>.02</td>
<td>.09</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td>.10</td>
<td>.36</td>
<td>.20</td>
<td>.21</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td>.30</td>
<td>.41</td>
<td>.26</td>
<td>.28</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td>.60</td>
<td>.60</td>
<td>.50</td>
<td>.54</td>
</tr>
<tr>
<td></td>
<td>.60</td>
<td>.10</td>
<td>.76</td>
<td>.51</td>
<td>.64</td>
</tr>
<tr>
<td></td>
<td>.60</td>
<td>.30</td>
<td>.87</td>
<td>.66</td>
<td>.77</td>
</tr>
<tr>
<td></td>
<td>.60</td>
<td>.60</td>
<td>.99</td>
<td>.95</td>
<td>.99</td>
</tr>
<tr>
<td>50</td>
<td>.10</td>
<td>.10</td>
<td>.20</td>
<td>.01</td>
<td>.08</td>
</tr>
<tr>
<td></td>
<td>.10</td>
<td>.30</td>
<td>.23</td>
<td>.02</td>
<td>.10</td>
</tr>
<tr>
<td></td>
<td>.10</td>
<td>.60</td>
<td>.30</td>
<td>.04</td>
<td>.18</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td>.10</td>
<td>.68</td>
<td>.29</td>
<td>.52</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td>.30</td>
<td>.76</td>
<td>.40</td>
<td>.65</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td>.60</td>
<td>.93</td>
<td>.70</td>
<td>.88</td>
</tr>
<tr>
<td></td>
<td>.60</td>
<td>.10</td>
<td>.98</td>
<td>.94</td>
<td>.98</td>
</tr>
<tr>
<td></td>
<td>.60</td>
<td>.30</td>
<td>1.00</td>
<td>.98</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>.60</td>
<td>.60</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

*Note.* W is E. J. Williams’ (1959) method. D1 and D2 are Methods D1 and D2, discussed in the present article. ES = effect size.
seem to be of potential interest for solving the problem at hand. However, Methods D1 and D2 do not readily extend to the problem of comparing these dependent correlations, and we leave to future researchers the issue of how best to proceed.

**AUTHOR NOTES**

Rand R. Wilcox is a professor of psychology at the University of Southern California in Los Angeles. His interests include quantitative methods, particularly robust techniques. Tian Tian is a graduate student at the University of Southern California. Her specialty is quantitative methods.

**REFERENCES**


*Manuscript submitted January 18, 2007
Revision accepted for publication April 20, 2007*