Comparing Correlated Correlation Coefficients

Xiao-Li Meng
Department of Statistics, University of Chicago

Robert Rosenthal
Harvard University

Donald B. Rubin
Department of Statistics, Harvard University

The purpose of this article is to provide simple but accurate methods for comparing correlation coefficients between a dependent variable and a set of independent variables. The methods are simple extensions of Dunn & Clark's (1969) work using the Fisher z transformation and include a test and confidence interval for comparing two correlated correlations, a test for heterogeneity, and a test and confidence interval for a contrast among k (>2) correlated correlations. Also briefly discussed is why the traditional Hotelling's t test for comparing correlated correlations is generally not appropriate in practice.

Often in psychological research we want to compare two correlations obtained from a single sample of subjects where each correlation is between a predictor variable (X_1 or X_2) and a single common dependent variable (Y). For example, a clinical research worker may want to compare the correlations of a long expensive test (e.g., the Wechsler Adult Intelligence Scale) and a short, inexpensive test (e.g., the Shipley-Institute of Living Scale for Measuring Intellectual Impairment) with a dependent variable (e.g., gains from educational therapy). If the two predictor variables do not differ significantly in their ability to predict the outcome and if the absolute magnitudes of their correlations are similar, the shorter test may be used with economic advantage. The traditional test for the significance of the difference between these correlated correlation coefficients has been Hotelling's t test (Hotelling, 1940; cited by, e.g., Guilford & Fruchter, 1978; McNemar, 1969; Walker & Lev, 1953), which is often regarded as "exact."

However, Steiger (1980), in his comprehensive review of the literature on the comparison of correlated correlation coefficients, has alerted the psychological research community to some potentially serious problems with the Hotelling method. In fact, such warnings can be found in statistical literature as early as in Hotelling's own article: "The advantages of exactness and of freedom from the somewhat special trivariate normal assumption are attained *at the expense of sacrificing the precise applicability of the results to other sets of the predictors*" (Hotelling, 1940, pp. 276–277). Williams (1959) proposed a modified t test to overcome this difficulty and emphasized that the limitations of Hotelling's t test should be noted in practice. But apparently, even after Steiger's (1980) explicit warning, "[Hotelling's t test] need not and should not be used for this purpose", (p. 246), Hotelling's t test is still the standard test for comparing two correlated correlation coefficients in psychological research, probably because of misunderstandings concerning the exactness and the apparent simplicity of it relative to alternatives. In the Appendix, after deriving our results, we use a simple example to emphasize the inappropriateness of Hotelling's t test for comparing two correlated correlation coefficients and to show the unusual sense in which it is exact. Our results provide generally valid tests and confidence intervals for comparing two or more correlated correlations.

Our Results

It is well-known that Fisher's z transformation of sample correlation coefficients improves the normality substantially, especially for small sample sizes and extreme sample correlations. An asymptotic test for comparing two correlated correlation coefficients using Fisher's z transformation was proposed in Dunn and Clark (1969, 1971), and its superior performance was confirmed in several studies (e.g., Neill & Dunn, 1975; Steiger, 1980). Thus it should be preferred to Williams's (1959) modified t test, which is on the original r scale. Here, we first present a simple procedure for comparing two correlated correlation coefficients, which is equivalent to Dunn & Clark's test asymptotically but is in a rather simple and thus easy-to-use form. Then we give a simple generalization of this procedure to any number of correlation coefficients having one variable in common, so that we can test the heterogeneity of a set of correlated correlation coefficients. We also provide a generalization of this procedure, so that we may ask focused theoretical questions about the ordering of the elements of the set of correlated correlations; for example, testing whether these correlations with a common variable follow the pattern of magnitudes that a particular theory would predict.

We determined the order of authorship alphabetically. The work was partially supported by National Science Foundation Grant SES-88-05433 and by the Spencer Foundation, although the views expressed are our responsibility.

We wish to thank three anonymous reviewers for very helpful comments.

Correspondence concerning this article should be sent to Robert Rosenthal, Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, Massachusetts 02138.

Psychological Bulletin, 1992, Vol. 111, No. 1, 172–175
 0033-2909/92/$3.00

Comparing Two Correlated Correlations

The following equation yields a Z (normal curve) test for the significance of the difference between two sample correlation coefficients r_{YX_1} and r_{YX_2} where variables X_1 and X_2 are predictors of dependent variable Y:

$$Z = (z_{r_1} - z_{r_2})\sqrt{\frac{N-3}{2(1-r_x)h}}, \tag{1}$$

where N is the number of subjects, z_{r_i} is the Fisher z-transformed $r_i \equiv r_{YX_i}$, r_x is the correlation between the two predictor variables X_1 and X_2 (i.e., $r_{X_1X_2}$),

$$h = \frac{1-f\bar{r}^2}{1-\bar{r}^2} = 1 + \frac{\bar{r}^2}{1-\bar{r}^2}(1-f), \tag{2}$$

$$f = \frac{1-r_x}{2(1-\bar{r}^2)}, \text{ which must be } \leq 1, \tag{3}$$

and $\bar{r}^2$ is the mean of the r_i^2, i.e., $(r_1^2 + r_2^2)/2$. The bound on f is derived from constraints among the correlation coefficients (i.e., the covariance matrix must be nonnegative), and f should be set to 1 if $(1-r_x)/(2(1-\bar{r}^2)) > 1$. Confidence intervals for the difference in $z'_i s$ can be obtained from Equation 1. For example, a 95% confidence interval is given by

$$z_{r_1} - z_{r_2} \pm 1.96\sqrt{\frac{2(1-r_x)h}{N-3}}. \tag{4}$$

Example 1

Fifteen psychological experimenters were filmed as they conducted a standard psychological experiment on person perception (Rosenthal, 1976). Later, it was possible to measure for all experimenters the degree to which they obtained data from their subjects consistent with an expectation that had been experimentally induced. Observers of the film rated each experimenter on the degree of professionalism and the degree of friendliness shown to subjects during the experiment. The correlation between professionalness of manner and the subsequent degree of experimenter expectancy effect (r_1) was .63, that between friendliness and expectancy effect (r_2) was −.03, and that between friendliness and professionalness (r_x) was −.19. We want to compare $r_1 = .63$ with $r_2 = -.03$. Here, we have $N = 15$; $z_{r_1} = .741$ and $z_{r_2} = -.030$; $r_x = -.19$; $\bar{r}^2 = [(.63)^2 + (-.03)^2]/2 = .1989$; and $f = (1 - (-.19))/2(1 - .1989) = .7427$, so that $h = (1 - f\bar{r}^2)/(1 - \bar{r}^2) = 1.0639$. Thus

$$Z = 0.771 \times \sqrt{\frac{12}{2 \times 1.19 \times 1.0639}} = 1.68,$$

$p = .047$, one-tailed.

Similarly, a 95% confidence interval (two-tailed) for $z_{r_1} - z_{r_2}$ is

$$0.771 \pm 1.96\sqrt{\frac{2 \times 1.19 \times 1.0639}{12}} = (-0.129, 1.671).$$

Testing the Heterogeneity of a Set of Correlated Correlations

If there are more than two predictor variables, so that there are more than two correlations between predictors and a common dependent variable, we can readily test the significance of their heterogeneity by means of a simple extension of Equation 1 yielding a χ^2 test:

$$\chi^2(k-1) = \frac{(N-3)\sum_i (z_{r_i} - \bar{z}_r)^2}{(1-r_x)h}, \tag{5}$$

where $\bar{z}_r$ is the mean of the z_{r_i}. The resulting χ^2 statistic is distributed on $k-1$ degrees of freedom where k is the number of predictive correlations being tested for heterogeneity. In this equation, r_i takes on $k \geq 2$ values, so that in the definition of h given by Equations 2 and 3, $\bar{r}^2$ is the average of all k values of r_i^2, and r_x is the *median* intercorrelation among the predictor variables being tested for heterogeneity.

Example 2

Table 1 shows the data of Example 1 augmented by two other variables describing experimenters' behavior during the conduct of the experiment. We want to know the degree to which the four predictor–criterion correlations differ significantly among themselves, namely, do the r_is of .63, .53, .54, and −.03 differ significantly? In this case, we have $N = 15$; $z_{r_i} = .741, .590, .604$, and $-.030$; $\bar{z}_r = .47625$; $r_x = .37$, $\bar{r}^2 = [(.63)^2 + (.53)^2 + (.54)^2 + (-.03)^2]/4 = .2426$; $f = (1 - .37)/2(1 - .2426) = .4159$, so $h = (1 - f\bar{r}^2)/(1 - \bar{r}^2) = 1.1871$. Thus,

$$\chi^2(3) = \frac{(12)(0.3556)}{(0.63)(1.1871)} = 5.71, \quad p = 0.127.$$

Testing a Contrast Among Correlated Correlation Coefficients

Our theory may call for a test of the hypothesis that certain predictors will do a better job than others in predicting the criterion variable. Contrasts in general allow us to ask focused questions of the data, so that precise tests of hypotheses are possible (Rosenthal & Rosnow, 1985). It is a simple matter to test contrasts among a set of correlated correlations by means of the following Z test,

$$Z = \sum \lambda_i z_{r_i} \sqrt{\frac{N-3}{(\sum \lambda_i^2)(1-r_x)h}}, \tag{6}$$

where the λ_is are the contrast weights assigned to each of the z_{r_i}s to be tested and all other terms are as defined earlier. (When k is large and some z_{r_i}s have zero λ_i for some contrasts, more accurate tests are obtained by using the local values of h and r_x

Table 1
Correlations Between Experimenter Behavior and Expectancy Effects ($N = 15$)

Behavior	Expectancy effect (Y)	Professional (A)	Dominant (B)	Likable (C)
Professional	.63	—		
Dominant	.53	.38	—	
Likable	.54	.36	.38	—
Friendly	−.03	−.19	.12	.60

relevant to the z_{r_i}s with nonzero λ_is rather than the global values that are based on all k parameters.) Confidence intervals for the contrast can be obtained from Equation 6. For example, a 95% confidence interval is given by

$$\sum \lambda_i z_{r_i} \pm 1.96 \sqrt{\frac{(\sum \lambda_i^2)(1 - r_x)h}{N - 3}}. \quad (7)$$

Example 3

Assume that our theory states that the variable friendly of Table 1 should predict expectancy effects less well than the average prediction obtained from the other three variables: professional, dominant, and likable. We should select our λ_is (which, as in the case of all contrast weights, must sum to zero) as -3, 1, 1, 1, respectively. Then in Equation 6, $\sum \lambda_i z_{r_i} = (-3)(-.030) + (1)(.741) + (1)(.590) + (1)(.604) = 2.025$ and $\sum \lambda_i^2 = (-3)^2 + (1)^2 + (1)^2 + (1)^2 = 12$, and therefore

$$Z = 2.025 \sqrt{\frac{12}{(12)(0.63)(1.1871)}} = 2.34,$$

$p = 0.01$, one-tailed.

An alternative, even simpler, way to compute a contrast is available once we have computed the overall test for heterogeneity. We simply compute the correlation coefficient between the λ_is of our contrast weights and their corresponding z_{r_i}s (i.e., $r_{\lambda z_r}$), then multiply it by the square root of the overall $\chi^2(k-1)$. That is,

$$Z = r_{\lambda z_r} \sqrt{\chi^2(k - 1)}, \quad (8)$$

where for our example $r_{\lambda z_r} = .9802$ and $\chi^2(k - 1) = 5.71$ from Example 2. Therefore,

$$Z = (0.9802)\sqrt{(5.71)} = 2.34,$$

which is identical to the Z obtained before. Similarly, a 95% confidence interval (two-tailed) for z_{r_1} versus the average of z_{r_2}, z_{r_3} and z_{r_4} is

$$2.025 \pm 1.96 \sqrt{\frac{(12)(0.63)(1.1871)}{12}} = (0.330, 3.720).$$

References

Dunn, O. J., & Clark, V. A. (1969). Correlation coefficients measured on the same individuals. *Journal of the American Statistical Association, 64,* 366–377.

Dunn, O. J., & Clark, V. A. (1971). Comparison of tests of the equality of dependent correlation coefficients. *Journal of the American Statistical Association, 66,* 904–908.

Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education* (6th ed.). NY: McGraw-Hill.

Hotelling, H. (1940). The selection of variates for use in prediction, with some comments on the general problem of nuisance parameters. *Annals of Mathematical Statistics, 11,* 271–283.

McNemar, Q. (1969). *Psychological statistics* (4th ed.). New York: Wiley.

Neill, J. J., & Dunn, O. J. (1975). Equality of dependent correlation coefficients. *Biometrics, 31,* 531–543.

Rosenthal, R. (1976). *Experimenter effects in behavioral research* (Rev. ed.). New York: Irvington.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis.* New York: Cambridge University Press.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245–251.

Walker, H. M., & Lev, J. (1953). *Statistical inference.* New York: Holt.

Williams, E. J. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society, Series B, 21,* 396–399.

Appendix

Technical Discussion

Assume that the joint distribution of the dependent variable Y and the k predictor variables $X_1, \ldots, X_k$ is normal, with common correlation R_x among the predictors (this condition does not appear to be critical in practice) and correlations $R_1, \ldots, R_k$ between the predictors and Y. Let r_i be the sample correlation between Y and X_i, z_{r_i} be the corresponding Fisher's z transformation of r_i,

$$z_{r_i} = \frac{1}{2} ln\left(\frac{1 + r_i}{1 - r_i}\right),$$

and z_{R_i} be the corresponding Fisher's z transformation of $R_i (i = 1, \ldots, k)$. Standard large sample theory shows that

$$\sqrt{N - 3}(z_{r_i} - z_{R_i})$$

has a standard (mean zero, variance one) normal distribution where the correlation between z_{r_i} and $z_{r_j} (i \neq j)$ is

$$\text{corr}\,(z_{r_i}, z_{r_j}) = R_x - \frac{R_i R_j}{2} + \frac{R_i R_j (R_i R_j - R_x)^2}{2(1 - R_i^2)(1 - R_j^2)}; \quad \text{(A1)}$$

see, for example, Steiger (1980).

Because Equation 1 is a special case of Equation 6 when $k = 2$, and because Equation 5 follows from Equation 1 and standard results on normal distributions, we only need to justify Equation 6 with $\Sigma\lambda_i = 0$. We wish to simplify

$$\text{var}\,(\sum_i \lambda_i z_{r_i}) = \left[\sum_i \lambda_i^2 + \sum_{i \neq j} \lambda_i \lambda_j \,\text{corr}\,(z_{r_i}, z_{r_j})\right]/(N - 3) \quad \text{(A2)}$$

by substituting a common estimate for all $\text{corr}(z_{r_i}, z_{r_j})$, say c, to obtain

$$\text{var}\,(\sum_i \lambda_i z_{r_i}) \approx (1 - c) \sum_i \lambda_i^2/(N - 3). \quad \text{(A3)}$$

Accepting Approximation A3, the issue is how to estimate c. Suppose we substitute the median r_x for R_x and each then substitute r_i as a possible estimate of both R_i and R_j in $\text{corr}(z_{r_i}, z_{r_j})$ to obtain k possible values of $\text{corr}(z_{r_i}, z_{r_j})$. The average of these values, assuming the r_i^2s are either all small or not small but nearly equal, is

$$c = 1 - (1 - r_x)h, \quad \text{(A4)}$$

where h is defined by Equations 2 and 3 with $\overline{r^2}$ the average of all r_i^2, and so we obtain Equation 6.

Although the approximation given by Equations A3 and A4 is based on a small r_i^2 or not too variable r_i^2 assumption, it appears to work quite well for a large range of r_is, as we now illustrate. Using the data of Example 3, we calculated the correct expression (A2) using (A1) and the approximation (A3) using (A4). Table A1 contains the results. The approximation works very well in this example even though the r_is are not small at all (three r_is are bigger than .5) and the range is quite wide (from −.03 to .63). In the presence of extreme rs (e.g., $|r| \geq 0.8$), which is quite rare in psychological research, the correct expression (A2) for the large sample variance would be preferred, although the approximation (A3) can still be used for calculating Equation 6 as a quick screening test. Under the null hypothesis that $R_i = R_j$ for all i, j, one can legitimately substitute $\bar{r}$ for each r_i and obtain a modified expression with $(\bar{r})^2$ in place of $\overline{r^2}$. This substitution is only correct strictly under the null hypothesis of all R_i equal, whereas our use of $\overline{r^2}$ is appropriate more generally, that is, for confidence intervals and tests of contrasts; see the third line of Table A1 for an example.

Hotelling's (1940) test is nearly always inappropriate because it fixes the Xs and tests a different null in general. To take an extreme example slightly modified from Steiger (1980), suppose X_1 and X_2 are independent, mean zero, variance one, normal variables, and $Y = (0.5)^{1/2}(X_1 + X_2)$. Hotelling's test essentially always rejects despite the fact that $R_1 = R_2 = (0.5)^{1/2}$, but this is not an error in Hotelling's test. It is correctly rejecting its own null hypothesis: $R_1(s_1/\sigma_1) = R_2(s_2/\sigma_2)$, where s_i is the standard deviation calculated from the set of fixed X_i, and σ_i is the population standard deviation of $X_i (i = 1, 2)$ (Dunn & Clark, 1969). In other words, Hotelling's test is exact but only tests the null that we are interested in, that is, $R_1 = R_2$, when the sample variance of X_i equals the population variance of X_i for $i = 1, 2$, which clearly almost never occurs with any real data where X and X_2 are random variables. The Hotelling's null hypothesis is certainly almost never of substantive interest.

Table A1
Comparison of Approximations

Procedure	$v = V(\Sigma\lambda_i z_{r_i})$	$s = \sqrt{v}$	$Z = \frac{\Sigma\lambda_i z_{r_i}}{s}$	$p = 1 - \Phi(Z)$
Exact calculation (Expression A2)	0.7482	0.8650	2.3419	.0096
Approximation (Expression A3)	0.7478	0.8648	2.3423	.0096
Expression A3 with $(\bar{r})^2$ in place of $\overline{r^2}$	0.7123	0.8440	2.3994	.0082

Received July 25, 1990
Accepted July 18, 1991 ■