JOINT DISTRIBUTION OF SOME INDICES BASED ON CORRELATION COEFFICIENTS

BY

LARRY V. HEDGES and INGRAM OLKIN

TECHNICAL REPORT NO. 188
MARCH 1983

PREPARED UNDER THE AUSPICES OF
NATIONAL SCIENCE FOUNDATION GRANT
MCS 81-04262
Ingram Olkin, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
JOINT DISTRIBUTION OF SOME INDICES BASED
ON CORRELATION COEFFICIENTS

by

Larry V. Hedges and Ingram Olkin

TECHNICAL REPORT NO. 188
MARCH 1983

Prepared Under the Auspices
of
National Science Foundation Grant
MCS 81-04262
Ingram Olkin, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
JOINT DISTRIBUTION OF SOME INDICES
BASED ON CORRELATION COEFFICIENTS

Correlation coefficients and other indices derived from correlation coefficients have been used extensively in psychological research for most of this century. The paucity of natural scales of measurement (non-arbitrary scale factors) for many areas of psychology continues to make scale-free measures of association important in behavioral science. This is especially true in areas of psychology that deal with personality and mental ability factors. In these areas the search for absolute scales of measurement has been all but abandoned and thus scale-free measures of association remain important if not essential.

The asymptotic distribution of the sample product-moment correlation was known to Pearson and Filon (1898) and the exact sampling distribution was obtained by Fisher (1915).

All but the simplest data analyses involve more than two variables. Most data analytic problems require two natural generalizations of the bivariate correlation: the partial and multiple correlations. Fisher (1924) obtained the distribution of the partial correlation coefficient, which is that of the product moment correlation with reduced degrees of freedom. Fisher (1928) also obtained the distribution of the squared multiple correlation. Other generalizations of the bivariate correlation coefficient have sometimes been suggested, such as the "partial-multiple correlation" in which one set of variates are used to predict a criterion variable while controlling for the values of a third set of variables (see e.g., Rao, 1973, p. 268). The exact distribution of the partial-
multiple correlation was obtained by Das Gupta (1977).

Despite the importance of these correlation coefficients as statistical tools in behavioral science research, there has been relatively little work on distribution theory for combinations of indices. For each of the partial, multiple, and partial-multiple correlations, marginal distributions are known which yields tests that the population parameter is zero. Yet in each case, the joint distribution of two or more of the correlations is unknown. One of the reasons for the lack of joint distribution theory is the complexity of the marginal distributions. For many problems, the mathematics involved in obtaining the exact joint distribution appears intractable. The asymptotic distributions are quite tractable, however. Olkin and Siotani (1976) obtained the asymptotic distribution of many functions of sample correlation matrices. For example they obtained the asymptotic joint distribution of the determinants of a correlation matrix and its principal submatrices. Hedges and Olkin (1981) extend the results of Olkin and Siotani to include the asymptotic joint distribution of certain differences between squared multiple correlations.

The present paper extends previous results by presenting the asymptotic joint distribution of arbitrary sets of partial, multiple, or partial-multiple correlations. These results can be used, for example, to obtain an asymptotic confidence interval for changes in squared multiple correlations or for the difference between partial correlations that control for different variables. Section 2 contains a general theorem on the asymptotic joint distribution of the determinants of arbitrary correlation matrices of variables. This theorem is the fundamental tool used to obtain the joint
distributions of partial, multiple, and partial-multiple correlations given in Section 3. Some special results for the trivariate normal distribution are given in Section 4. Section 5 contains examples of the application of methods described in this paper.
2. **A Fundamental Theorem.**

Determinants of correlation matrices are involved in the definitions of many indices based on correlations. Because correlation matrices of normal variates are functions of sample moments, they will have asymptotic normal distributions. Consequently, the distribution of the determinants of correlation matrices is therefore a starting point for obtaining the distributions of functions of such determinants.

In some cases the asymptotic covariance matrix has a simple, compact form. In others, the expressions are quite complicated. Rather than present very complicated expressions, we indicate how they may be determined numerically.

Consider a partition of \( p+1 \) variables into \( k+1 \) vectors \( x_0, x_1, \ldots, x_k \), where \( x_i \) is a \( p_i \)-dimensional column vector, \( i = 0, 1, \ldots, k \) with \( p_0 + p_1 + \cdots + p_k = p+1 \) and where \( p_0 = p_1 = 1 \). Further, we start with a sample of size \( n \) from a \((p+1)\)-variate normal distribution, where the variables are subdivided as above.

We require a notation to denote a correlation matrix made up of a subset of the vectors. The symbols \( R(s_1, s_2, \ldots, s_m) \) and \( P(s_1, s_2, \ldots, s_m) \) denote sample and population correlation matrices, respectively, made up of vectors \( x_{s_1}, \ldots, x_{s_m} \). In some instances we let \( S = \{s_1, \ldots, s_m\} \) and write simply \( R(S) \).

Denote by \( d \) and \( \delta \) the vectors of determinants of sample and population correlation matrices listed in lexicographic order, i.e.,

\[
d = (|R(0)|, |R(1)|, \ldots, |R(0,1)|, \ldots, |R(0,1,\ldots, )|')',
\]
\[
\delta = (|P(0)|, |P(1)|, \ldots, |P(0,1)|, \ldots, |P(0,1,\ldots, )|')'.
\]
In this notation, \( R(0,1,\ldots,k) \) represents the \( p+1 \times p+1 \) matrix of sample correlations.

**Theorem 1.** For a sample of \( n \) observations from a \((p+1)\)-variate normal distribution, the asymptotic distribution of \( d \) is given as

\[
\sqrt{n} (d-\delta) \sim \mathcal{N}(0, \Psi_{\infty}),
\]

where the covariance matrix of the limiting distribution has elements of the form

\[
\text{Cov}_{\infty}(\mathcal{R}(U), \mathcal{R}(V)) = 2 |\mathcal{P}(U)||\mathcal{P}(V)| \sum_{i,j \in U} \sum_{\lambda, \mu \in V} \rho_{ij} \rho_{\lambda \mu} \left\{ \frac{1}{2} \rho_{ij}^2 \rho_{\lambda \mu} + \rho_{ij}^2 \rho_{\lambda \mu} + \rho_{ij} \rho_{\lambda \mu} \right\}.
\]

where \( \rho_{ij} \) and \( \rho_{\lambda \mu} \) are the elements of \( \mathcal{P}^{-1}(U) \) and \( \mathcal{P}^{-1}(V) \), respectively.

**Proof.** Denote the vector of all sample and population correlations from a \((p+1)\)-variate normal distribution (in lexicographic order) by

\[
r = (r_{01}, r_{02}, \ldots, r_{p-1,p}),
\]

\[
\rho = (\rho_{01}, \rho_{02}, \ldots, \rho_{p-1,p}).
\]

The vector \( r \) as a function of sample moments has an asymptotic multivariate normal distribution i.e.,

\[
\sqrt{n} (r-\rho) \sim \mathcal{N}(0, \Phi_{\infty}).
\]

The asymptotic covariance matrix \( \Phi_{\infty} \) was first obtained by Pearson and Filon (1898). In our notation \( \Phi_{\infty} = \text{cov}_{\infty}[r_{ij}, r_{\lambda \mu}] \), where
\[
\text{Cov}_\omega(r_{ij},r_{lm}) = \frac{1}{2} \rho_{ij} \rho_{km} (\rho_{ik}^2 + \rho_{im}^2 + \rho_{jl}^2 + \rho_{jm}^2) + \rho_{il} \rho_{jm}
\]
\[
+ \rho_{im} \rho_{jl} - (\rho_{ij} \rho_{il} \rho_{im} \rho_{jl} + \rho_{ij} \rho_{jl} \rho_{jm} + \rho_{il} \rho_{lj} \rho_{jm} + \rho_{im} \rho_{mj} \rho_{ml}).
\]

It is well-known that if \( T = (T_1, \ldots, T_k)' \) has an asymptotic \( k \)-variate normal distribution given by \( \sqrt{n} (T-\theta) \sim N(0,\Sigma) \), and if \( f_1, \ldots, f_m \) (\( m \leq k \)) are functions of \( \theta \) having first and second derivatives in a neighborhood of \( \theta \), then \( f(T) \equiv [f_1(T), \ldots, f_m(T)] \) has an asymptotic \( m \)-variate normal distribution given by

\[
\sqrt{n} [f_1(T)-f_1(\theta), \ldots, f_m(T)-f_m(\theta)]' \sim N(0, \Lambda' \Psi \Lambda),
\]

where \( \Lambda = (\lambda_{ij}) \) and \( \lambda_{ij} = \partial f_i(t)/\partial \theta_j \) evaluated at \( t = \theta \). An application of this theorem with \( T = r, \theta = \rho, f_1(r) = |R(0)|, \ldots, f_m(p) = |R(0, \ldots, k)| \) and the fact that

\[
\frac{\partial R(U)}{\partial r_{ij}} = 2|R(U)|r_{ij} \quad \text{if } r_{ij} \text{ is an element of } R(U),
\]

and

\[
\frac{\partial R(U)}{\partial r_{ij}} = 0 \quad \text{if } r_{ij} \text{ is not an element of } R(U)
\]
yields the result. ||

Remark. Note that the theorem above remains true if \( d \) is replaced with a vector of arbitrary nonsingular matrices whose elements are correlation coefficients.


We now use Theorem 1 to obtain the asymptotic joint distributions of a vector of squared multiple correlations (Section 3.1), partial correlations
(Section 3.2), and partial-multiple correlations (Section 3.3), and the distribution of Coleman's generalized "standardized regression coefficients" (Section 3.4).

3.1 Squared Multiple Correlations.

Denote the vector of sample and population squared multiple correlations, arranged in lexicographic order, by

\[ \bar{r} = (r_{01}^2, r_{02}^2, \ldots, r_{0k}^2, r_{0(12)}^2, \ldots, r_{0(12\cdots k)}^2) , \]

\[ \bar{\rho} = (\rho_{01}^2, \rho_{02}^2, \ldots, \rho_{0k}^2, \rho_{0(12)}^2, \ldots, \rho_{0(12\cdots k)}^2) , \]

where \( r_{0(s_1, \ldots, s_m)} \) denotes the sample multiple correlation between \( X_0 \) and \( X_{s_1}, \ldots, X_{s_m} \). Recall that \( X_{i}, i = 1, \ldots, k \) is itself a vector.

**Theorem 2.** The asymptotic distribution of \( \bar{r} \) is given by

\[ \sqrt{n} (\bar{r} - \bar{\rho}) \sim h(0, \Gamma_{\infty}) , \]

where the covariance matrix of the limiting distribution has elements of the form

\[ \text{Cov}(r_{0(\alpha)}^2, r_{0(\alpha)}^2) = 4 \rho_{0(\alpha)}^2 [1 - \rho_{0(\alpha)}^2]^2 , \]

\[ \text{Cov}(r_{0(\alpha)}^2, r_{0(\beta)}^2) = \text{Cov}(|R(\alpha)|, |R(\beta)|) \]

\[ + \frac{|P(0,\alpha)| |P(0,\beta)|}{|P(\alpha)| |P(\beta)|} \text{Cov}(|R(0,\alpha)|, |R(0,\beta)|) \]

\[ - \frac{|P(0,\alpha)|}{|P(\alpha)|^2 |P(\beta)|} \text{Cov}(|R(0,\alpha)|, |R(\beta)|) - \frac{|P(0,\beta)|}{|P(\alpha)| |P(\beta)|^2} \text{Cov}(|R(\alpha)|, |R(0,\beta)|) , \]
where \( \alpha \) and \( \beta \) denote sets of subscripts. The relevant covariance terms are given in Theorem 1.

**Proof.** The proof follows from a direct application of Olkin and Siotani (1976, Theorem 2.1) after writing

\[
\rho_{0(\alpha)}^2 = 1 - \frac{|R(0, \alpha)|}{|R(\alpha)|}. ||
\]

If \( C \) is a matrix of constants of order commensurate with \( \tilde{r} \), then

\[
\sqrt{n} (C\tilde{r} - C\tilde{\rho}) \sim \mathcal{N}(0, C\Gamma C').
\]

### 3.2 Partial Correlations.

Denote the vector of sample and population partial correlations between \( X_0 \) and \( X_1 \) for fixed \( X_2, X_3, \ldots, X_k \) (arranged in lexicographic order) by

\[
\rho^* = (r_{01(2)}, r_{01(3)}, \ldots, r_{01(23)}, \ldots, r_{01(2, \ldots, k)}'),
\]

\[
\rho^* = (\rho_{01(2)}, \rho_{01(3)}, \ldots, \rho_{01(23)}, \ldots, \rho_{01(2, \ldots, k)}').
\]

**Theorem 3.** The asymptotic distribution of \( \rho^* \) is

\[
\sqrt{n} (\rho^* - \rho^*) \sim \mathcal{N}(0, \Phi),
\]

where the covariance matrix of the limiting distribution has elements of the form
\[ \text{Cov}(r_{01}(\alpha), r_{01}(\beta)) = (1 - \rho_{01}^2(\alpha))^2, \]

\[ \text{Cov}(r_{01}(\alpha), r_{01}(\beta)) = \frac{1}{\sqrt{|P(0,\alpha)||P(0,\beta)||P(1,\alpha)||P(1,\beta)|}} \times \]

\[ \left\{ \text{Cov}(C_{01}(\alpha), C_{01}(\beta)) - \frac{\Gamma_{01}(\alpha)}{2|P(0,\alpha)|} \text{Cov}(|R(0,\alpha)|, C_{01}(\beta)) - \frac{\Gamma_{01}(\beta)}{2|P(0,\beta)|} \text{Cov}(|R(0,\beta)|, C_{01}(\alpha)) \right. \]

\[ - \frac{\Gamma_{01}(\alpha)}{2|P(1,\alpha)|} \text{Cov}(|R(1,\alpha)|, C_{01}(\beta)) - \frac{\Gamma_{01}(\beta)}{2|P(1,\beta)|} \text{Cov}(|R(1,\beta)|, C_{01}(\alpha)) \]

\[ + \frac{\Gamma_{01}(\alpha)\Gamma_{01}(\beta)}{4 \sqrt{|P(0,\alpha)||P(0,\beta)||P(1,\alpha)||P(1,\beta)|}} \left( \frac{\text{Cov}(|R(0,\alpha)|, |R(0,\beta)|)}{|P(0,\alpha)||P(0,\beta)|} + \frac{\text{Cov}(|R(1,\alpha)|, |R(1,\beta)|)}{|P(1,\alpha)||P(1,\beta)|} \right) \]

\[ + \frac{\text{Cov}(|R(0,\alpha)|, |R(1,\beta)|)}{|P(0,\alpha)||P(1,\beta)|} + \frac{\text{Cov}(|R(1,\alpha)|, |R(0,\beta)|)}{|P(1,\alpha)||P(0,\beta)|} \right\}, \]

where \( C_{01}(\alpha) \) denotes the cofactor of \( r_{01} \) in the matrix \( R(0,1,\alpha) \),

\( \Gamma_{01}(\alpha) \) denotes the cofactor of \( \rho_{01} \) in the matrix \( P(0,1,\alpha) \).

Proof. the result follows by applying Theorem 2.1 of Olkin and Siotani (1976) in the representation

\[ r_{01}(\alpha)^2 = \frac{|R(0,1,\alpha)|}{\sqrt{|R(0,\alpha)||R(1,\alpha)|}}. \]

3.3 Partial-Multiple Correlations.

In this model we predict \( X_0 \) with variables \( X_1, \ldots, X_m \), and holding variables \( X_{m+1}, \ldots, X_k \) fixed. To simplify notation, let \( \alpha \) denote the subscripts \( 1, 2, \ldots, m \). Denote the vector of sample and population squared partial-multiple correlations, arranged in lexicographic order, by
\[ \tilde{r}_0(\alpha) = \left( r_0(\alpha) \cdot m+1, \ldots, r_0(\alpha) \cdot k, r_0(\alpha) \cdot m+2, \ldots, r_0(\alpha) \cdot m+1, \ldots, k \right), \]

\[ \tilde{\rho}_0(\alpha) = \left( \rho_0(\alpha) \cdot m+1, \ldots, \rho_0(\alpha) \cdot k, \rho_0(\alpha) \cdot m+2, \ldots, \rho_0(\alpha) \cdot m+1, \ldots, k \right). \]

**Theorem 4.** The asymptotic distribution of \( \tilde{r}_0(\alpha) \) is

\[ \sqrt{n} \left( \tilde{r}_0(\alpha) - \tilde{\rho}_0(\alpha) \right) \sim N(0, \Omega), \]

where the covariance matrix of the limiting distribution has elements of the form

\[ \text{Cov}(r_0^2(\alpha) \cdot \beta, r_0^2(\alpha) \cdot \gamma) = (a, b) \wedge (a, b)', \]

where

\[ a = \frac{P(\alpha)}{|P(\alpha)| |P(0, \alpha)|} \left( \begin{array}{c} 1/F(0, \beta) \cdot |P(\alpha)| \\ 1/F(0, \alpha) \cdot |P(\beta)| \end{array} \right), \]

\[ b = \frac{P(\alpha)}{|P(\alpha)| |P(0, \alpha)|} \left( \begin{array}{c} 1/F(0, \gamma) \cdot |P(\alpha)| \\ 1/F(0, \alpha) \cdot |P(\gamma)| \end{array} \right), \]

and \( \wedge \) is the matrix of covariances of

\[ |R(\beta)|, |R(0, \alpha, \beta)|, |R(\alpha, \beta)|, |R(0, \beta)|, |R(\gamma)|, |R(0, \alpha, \gamma)|, |R(\alpha, \gamma)|, |R(0, \gamma)| \]

as given by Theorem 1.

**Proof.** The result follows by an application of Theorem 2.1 of Olkin and Siotani (1976) to

\[ r_0^2(\alpha) \cdot \beta = \frac{|R(\beta)| |R(0, \alpha, \beta)|}{|R(\alpha)| |R(0, \beta)|}. \]
3.4 Coleman's Generalized Standardized Regression Coefficients

Many studies in education and sociology include a single well defined outcome variable, such as academic achievement, and a set of predictor variables that can be grouped into classes or "blocks." Often these blocks of variables are identified with common contexts, e.g., home background variables, school resource variables, or school configuration variables. Coleman (1975) suggested a generalization of the standardized regression coefficient as one means of describing the effects of one block of variables on the outcome while controlling for the effects of another block of variables. He defined the generalized standardized regression coefficients, and applied the technique to some data from the International Education Association (IEA) studies of school learning. We now provide some distribution theory for the indices that we defined.
4. **Special Results for the Trivariate Case.**

In the special case of three variables $X_0, X_1, X_2$ the results can be stated more simply: Define

$$g = (r_{01}^2, r_{12}^2, r_{02}^2)'$$

$$\gamma = (\rho_{01}^2, \rho_{02}^2, \rho_{12}^2)'$$

then

$$\sqrt{n} (g - \gamma) \sim N(0, \Sigma)$$

where

$$\sigma_{11} = 4\rho_{01}^2 (1-\rho_{01}^2)^2$$

$$\sigma_{22} = 4\rho_{12}^2 (1-\rho_{12}^2)^2$$

$$\sigma_{33} = 4\rho_{02}^2 (1-\rho_{02}^2)^2$$

$$\sigma_{12} = 4\rho_{01} \rho_{12} \text{Cov}(r_{01}, r_{12})$$

$$\sigma_{23} = 4\rho_{12} \left\{ \frac{\text{Cov}(r_{01}, r_{02})}{\rho_{01}} - (1-\rho_{12}^2) \right\}$$

$$\sigma_{13} = -4\rho_{01} \left\{ (1-\rho_{01}^2)^2 \frac{\rho_{01}}{\rho_{00}} + \text{Cov}(r_{01}, r_{02}) \frac{\rho_{02}}{\rho_{00}} + \text{Cov}(r_{01}, r_{12}) \frac{\rho_{12}}{\rho_{00}} \frac{\rho_{01}}{\rho_{00}} \right\}$$

Define

$$d = (r_{01}, r_{01.2})'$$

$$\delta = (\rho_{01}, \rho_{01.2})$$

then

$$\sqrt{n} (d - \delta) \sim N(0, \Sigma)$$

where

$$\sigma_{11} = (1-\rho_{01}^2)^2$$

$$\sigma_{22} = (1-\rho_{01.2}^2)^2$$

$$\sigma_{12} = \frac{1}{2} \rho_{01} \rho_{02} \rho_{12.0} \left( \frac{1-\rho_{01}^2 - \rho_{02}^2}{1-\rho_{12}^2} \right) + \frac{1}{2} \rho_{01} \rho_{12} \rho_{02.1} \left( \frac{1-\rho_{01}^2 - \rho_{02}^2}{1-\rho_{12}^2} \right)$$

$$+ \frac{(1-\rho_{01}^2)^2}{\sqrt{(1-\rho_{02}^2)(1-\rho_{12}^2)}}$$
Remark. The above results provide a correction to some typographical errors contained in Olkin and Siotani (1976).