

A LIKELIHOOD-RATIO-BASED NORMAL APPROXIMATION
FOR THE NON-NULL DISTRIBUTION
OF THE MULTIPLE CORRELATION COEFFICIENT

Panagis G. Moschopoulos and Govind S. Mudholkar
University of Texas at Dallas University of Rochester
Richardson, TX 75080 Rochester, NY 14260

Key Words and Phrases: hypergeometric functions; asymptotic expansions, cumulants of multiple correlation; Wilson and Hilferty transformation.

ABSTRACT

Let X_1, X_2, \dots, X_p be jointly distributed according to a multivariate normal distribution, and let ρ denote the multiple correlation coefficient between X_1 and X_2, X_3, \dots, X_p . Let $X_{1i}, \dots, X_{pi}, i = 1, \dots, N$, be a random sample from the distribution. The logarithm of the likelihood ratio statistic for testing the hypothesis that ρ is zero is $-(N/2)\log(1-R^2)$, where R is the sample multiple correlation coefficient. A Gaussian approximation to the non-null ($\rho \neq 0$) distribution of R is developed using the transformation $(T/E(T))^h$ where $T = -\log(1-R^2)$, and h is determined from the first three cumulants of T . The approximation is simple and accurate over a wide range of the parameters p, N , and ρ .

I. INTRODUCTION AND SUMMARY

Let R denote the sample multiple correlation coefficient between the first and remaining $(p-1)$ variables of a sample $X_{1i}, \dots, X_{pi}, i=1, \dots, N$ of size N from a p -variate normal population, and let ρ denote the corresponding population multiple correlation coefficient. The probability density function of R was obtained by Fisher in 1928. Since then most of the work on the subject has been devoted to facilitating computation of the

probabilities and percentiles of the distribution. Several approximations of the distribution of R are found in the literature (e.g., Khatri (1966), Gurland (1968), Lee (1971)). Such approximations are based on the fact (e.g., Hodgson (1965)) that $\bar{R}^2 = R^2 / (1 - R^2)$ admits the representation

$$\bar{R}^2 = \{(\tilde{\rho}\chi_{N-1} + Z)^2 + \chi_{p-2}^2\} / \chi_{N-p}^2, \quad (1.1)$$

where $\tilde{\rho}^2 = \rho^2 / (1 - \rho^2)$ and Z , χ_{N-1} , χ_{p-2}^2 and χ_{N-p}^2 are independently distributed standard normal, chi and chi-square variables respectively. The above relation is convenient for obtaining the characteristic function and consequently for constructing various moments-based approximations for the distribution of \bar{R}^2 . Lee (1971), has given an excellent account of such approximations in terms of central and non-central F-distributions; he also derived a number of normal approximations by applying Geary-Fieller (for example see Fieller, (1932)) reasoning to the power transformations of the numerator and denominator of (1.1).

From Lee's numerical evaluations it may be concluded that the quality of the noncentral-F approximation for R is good whereas that of the approximations based upon F and normal distribution is variable. It may also be noted that the Gaussian approximations of Lee are more suitable for obtaining probabilities rather than percentiles of R and often involve Edgeworth corrections.

In this paper we develop a simple normal approximation to the distribution of R using the likelihood ratio statistic for testing the hypothesis of zero multiple correlation in the population. In particular we use the multiple T of the log-likelihood ratio where

$$T = -\log(1 - R^2) \quad (1.2)$$

Our procedure is motivated from the fact that while the limiting null-distribution ($\rho=0$) of the multiple NT is a chi-square with $p-1$ d.f., the limiting distribution of T , under any fixed alternative $\rho \neq 0$, is normal. That is

$$\sqrt{N}(T + \log(1 - \rho^2)) / 4\rho^2 \xrightarrow{d} Z \quad (1.3)$$

as the sample size N tends to infinity. To accelerate the convergence to normality, we transform T to a random variable Y , where

$$Y = (T/(ET))^h \quad (1.4)$$

and choose h so that the skewness of the distribution of Y is reduced, thus obtaining an approximately symmetrizing transformation of T . The above approach is implicit in Wilson and Hilferty (1931) where a normal approximation for the cube root of a chi-square random variable is obtained. Since then it has been adopted by several authors. Jensen and Solomon (1972) employed (1.4) with a positive definite quadratic form in place of T . More recently, Mudholkar and Trivedi (1981) have used (1.4) to symmetrize the distribution of the variance of samples from non-normal populations.

In this paper we proceed as follows: In order to obtain the necessary cumulants of T we first derive the moment generating function (m.g.f.) of T . This is obtained in closed form in terms of a Gauss Hypergeometric function ${}_2F_1$ (c.f. (2.2) in the following section). Then, using results of Watson (1918) on the behavior of ${}_2F_1$ when some parameters are 'large' we expand the m.g.f. of T in a power series in $(N-1)^{-1}$. By taking the logarithm of the m.g.f. in that form, and expanding the logarithm, we obtain the cumulant generating function of T and hence the cumulants of T . The latter are expressed in power series in $(N-1)^{-1}$. This is accomplished in Section 2. We develop the normal approximation of the distribution of Y , and consequently of R , in Section 3. In Section 4 we report on the accuracy of the approximation based on a numerical evaluation consisting of a comparison between exact and approximate probabilities and percentiles (upper 1% and 5%) of the distribution of R . The approximation is shown to work satisfactorily over a wide range of parameter values and over the whole range of R . It is comparable in accuracy to Lee's (1971) central and non-central F approximations; moreover it is simple. In Section 5 we illustrate the use of the approximation in the construction of confidence intervals for ρ (of any confidence level).

2. THE CUMULANTS OF $T = -\log(1-R^2)$

It is well known that the density function of R^2 can be expressed as

$$f(R^2) = \frac{\Gamma(\frac{N-1}{2}) (1-\rho^2)^{\frac{N-1}{2}}}{\Gamma(\frac{p-1}{2}) \Gamma(\frac{N-p}{2})} (R^2)^{\frac{p-3}{2}} (1-R^2)^{\frac{N-p-2}{2}}$$

$$\cdot {}_2F_1\left(\frac{N-1}{2}, \frac{N-1}{2}; \frac{p-1}{2}; \rho^2 R^2\right), \quad (2.1)$$

where ${}_2F_1$ is the Gauss hypergeometric function defined by:

$${}_2F_1(a, b; c; z) = \sum_{j=0}^{\infty} \frac{\Gamma(a+j) \Gamma(b+j) \Gamma(c)}{\Gamma(a) \Gamma(b) \Gamma(c+j)} \frac{z^j}{j!}. \quad (2.2)$$

Using (2.2) and the relation (Luke Vol. I, 1969)

$${}_2F_1\left(\frac{n}{2}, \frac{n}{2}, \frac{n}{2} - t; \rho^2\right) = (1-\rho^2)^{-t} {}_2F_1\left(-t, -t; -t + \frac{n}{2}; \rho^2\right) \quad (2.3)$$

we obtain the moment generating function

$$M_n(t) = E e^{-t \log(1-R^2)} =$$

$$(1-\rho^2)^{-t} \frac{\Gamma(\frac{n}{2} - \frac{p-1}{2} - t)}{\Gamma(\frac{n}{2} - \frac{p-1}{2})} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n}{2} - t)} {}_2F_1\left(-t, -t; -t + \frac{n}{2}; \rho^2\right), \quad (2.4)$$

where $n = N-1$. The behavior of ${}_2F_1(a, b; c; z)$ for large parameters was studied by Watson (1918). From his results it follows that for fixed t and large n ,

$${}_2F_1\left(-t, -t; -t + \frac{n}{2}; \rho^2\right) = \frac{\Gamma(\frac{n}{2} - t)}{\Gamma(\frac{n}{2})} \left(\frac{n}{2}\right)^t \sum_{k=0}^{\infty} h_k(\rho^2) (-t)_k \left(\frac{n}{2}\right)^{-k}, \quad (2.5)$$

where $(a)_k = a(a+1) \dots (a+k-1)$ and the h_k 's which depend upon t, ρ^2 satisfy the relation:

$$\left(\frac{g(x)}{x}\right)^{-t-1} (1-z \cdot g(x))^t = \sum_{k=0}^{\infty} h_k(z) x^k, \quad (2.6)$$

where $g(x) = 1 - e^{-x}$. By a straightforward calculation Watson obtained

$$h_0(\rho^2) = 1 \text{ and } h_1(\rho^2) = \left(\frac{1}{2} - \rho^2\right)t + \frac{1}{2} \quad (2.7)$$

In the same manner we see that

$$h_2(\rho^2) = \left(\frac{\rho^4}{2} - \frac{\rho^2}{2} + \frac{1}{8}\right)t^2 + \left(-\frac{\rho^4}{2} + \frac{5}{24}\right)t + \frac{1}{12}, \quad (2.7a)$$

and
$$h_3(\rho^2) = b_3 t^3 + b_2 t^2 + b_1 t + \frac{1}{2},$$

where the b_i 's are given by

$$b_1 = -\frac{1}{3}\rho^6 + \frac{1}{4}\rho^4 + \frac{1}{48},$$

$$b_2 = \frac{1}{2}\rho^6 - \frac{1}{2}\rho^4 + \frac{1}{24}\rho^2 + \frac{1}{24}, \quad (2.8)$$

and
$$b_3 = -\frac{1}{6}\rho^6 + \frac{1}{4}\rho^4 - \frac{1}{8}\rho^2 + \frac{1}{48}.$$

Hence for the series in (2.5) we get the following asymptotic expansion:

$$\sum_{k=0}^{\infty} h_k(\rho^2) (-t)_k \left(\frac{n}{2}\right)^{-k} = 1 + \frac{a_1}{n} + \frac{a_2}{n^2} + \frac{a_3}{n^3}, \quad (2.9)$$

where a_1, a_2, a_3 are given by

$$\begin{aligned} a_1 &= (2\rho^2 - 1)t^2 - t, \\ a_2 &= (2\rho^4 - 2\rho^2 + \frac{1}{2})t^4 + (-4\rho^4 + 2\rho^2 + \frac{1}{3})t^3 \\ &\quad + (2\rho^4 - \frac{1}{2})t^2 - \frac{1}{3}t, \end{aligned} \quad (2.10)$$

and
$$a_3 = -8b_3 t^6 + (24b_3 - 8b_2)t^5 + (-16b_3 + 24b_2 - 8b_1)t^4 \\ + (-16b_2 + 24b_1)t^3 - 16b_1 t^2.$$

In order to use the form (2.4) of $M_n(t)$ for computing the cumulants of T we also need the following asymptotic expansion for the ratio of two gamma functions. (See Luke, 1969, p. 33).

$$\begin{aligned} \frac{\Gamma(z+a)}{\Gamma(z+b)} &= z^{a-b} \sum_{k=0}^{L-1} \frac{(-1)^k B^{(a-b+1)}(a)(b-a)_k}{k!} z^{-k} \\ &\quad + z^{a-b} O(z^{-L}), \end{aligned} \quad (2.11)$$

where B_k is the generalized Bernoulli polynomial of order k (Luke, 1969, page 19). Using (2.9) and (2.11), taking the logarithms of both sides in (2.4) and subsequently expanding we obtain after a lengthy but straightforward calculation, the cumulant generating function of T , $K_n(t) = -\log M_n(t)$ expanded to $O(n^{-3})$

$$K_n(t) = -t \log(1-\rho^2) + \sum_{r=1}^3 \frac{A_{r,p,\rho^2}(t)}{n^r} + C + O(n^{-4}) \quad (2.12)$$

Here C is independent of t and

$$A_{r,p,\rho^2}(t) = \Gamma_{r,p}(t) + F_{r,p}^2(t) \quad (2.13)$$

where Γ_r , $r = 1, 2, 3$ is the contribution to A_r coming from the expansion of the log-Gamma-ratio and F_r , $r = 1, 2, 3$ from the expansion of $\log({}_2F_1)$ in (2.4). Specifically we have the following:

$$\Gamma_{1,p}(t) = \gamma_1 \quad .$$

$$\Gamma_{2,p}(t) = \gamma_2 - \frac{1}{2} \gamma_1^2 \quad , \quad (2.14)$$

$$\Gamma_{3,p}(t) = \gamma_3 - \gamma_1 \gamma_2 + \frac{1}{3} \gamma_1^3$$

Here the γ 's are functions of t given by:

$$\gamma_1 = t^2 + pt \quad .$$

$$\gamma_2 = \frac{1}{6} \{ 3t^4 + (6v+10)t^3 + (3v^2+12v+9)t^2 + (3v^2+6v+2)t \} \quad ,$$

$$\begin{aligned} \gamma_3 = \frac{1}{6} \{ t^6 + (3v+7)t^5 + (3v^2+16v+17)t^4 + (v^3+16v^2+29v+17)t^3 \\ + (3v^3+15v^2 + 20v+6)t^2 + (2v^3+6v^2+4v)t \} \quad , \end{aligned} \quad (2.15)$$

where $v = p-1$, and with a_1, a_2, a_3 as defined in (2.10).

$$F_{1,p}^2(t) = a_1 \quad ,$$

$$F_{2,p}^2(t) = a_2 - \frac{1}{2} a_1^2 \quad ,$$

$$F_{3,p}^2(t) = a_3 - a_1 a_2 + \frac{1}{3} a_1^3 \quad . \quad (2.16)$$

Now differentiating $K_n(t)$ at $t=0$, we obtain the following expressions for the first four cumulants of T where we put back $N-1$ for n

$$\begin{aligned}
 k_1 &= -\log(1-\rho^2) + \frac{p-1}{N-1} + \frac{p^2-1}{2(N-1)^2} \\
 &+ \frac{(p-1)^3 + 3(p-1)^2 + 2(p-1)}{3(N-1)^3} + O((N-1)^{-4}) \\
 k_2 &= \frac{4\rho^2}{N-1} + \frac{4\rho^4 + 2(p-1)}{(N-1)^2} + \frac{2(p^2-1) + 8\rho^4((4/3)\cdot\rho^2-1)}{(N-1)^3} + O((N-1)^{-4}) \\
 k_3 &= \frac{24\rho^2(1-\rho^2)}{(N-1)^2} + \frac{8(p-1) + 96\rho^4(1-\rho^2)}{(N-1)^3} + O((N-1)^{-4}) \quad , \quad (2.17) \\
 k_4 &= \frac{160\rho^4(2\rho^2-3) + 192\rho^2}{(N-1)^3} + O((N-1)^{-4}) \quad .
 \end{aligned}$$

The exact form of the $O((N-1)^{-4})$ term in k_1 is $[(p-1)^4/4 + p(p-1)^2]/(N-1)^4$. Also ignoring terms involving powers of ρ , the $O((N-1)^{-4})$ terms corresponding to k_2 , k_3 and k_4 are $2(p-1)^3 + 6(p-1)^2 + 4(p-1)$, $12(p-1)(p+1)$ and $48(p-1)$ respectively. These terms are used in the calculations in the following sections.

Remark 1: It is well known that when $\rho = 0$,
 $-N \log(1-R^2) \xrightarrow{d} \chi^2_{p-1}$ as $N \rightarrow \infty$. An examination of the four cumulants in (2.17) shows that for large N and $\rho=0$, they agree with those of a central chi-square with $p-1$ d.f.

Remark 2: The mean and variance of the normal approximation for Y in section 3 are expressed in power series in $[(N-1)k_1]^{-1}$. Hence the accuracy of k_1 is important. It should be noted that the latter depends on both the sample size N and the relative magnitude of p with respect to N .

3. A TRANSFORMATION TO NORMALITY

As indicated earlier, for any fixed $\rho \neq 0$, the limiting distribution of $N(T + \log(1-\rho^2))$ is normal with mean zero and variance $4\rho^2$. Hence, by Mann-Wald theorem, any differentiable function of $T = -\log(1-R^2)$ is also asymptotically normally distributed. Here we consider the class of functions $\gamma = (T/k_1)^h$, and choose h as in

Jensen and Solomon (1972). Using a simple expansion, the expectation $\mu_1'(h)$ of Y is obtained as

$$\begin{aligned} \mu_1'(h) = & 1 + \frac{h(h-1)}{2[(N-1)k_1]} \phi_2 + \frac{h(h-1)(h-2)}{24[(N-1)k_1]^2} [4\phi_3 + 3(h-3)\phi_2^2] \\ & + O([(N-1)k_1]^{-3}), \end{aligned} \quad (3.1)$$

where $\phi_i = \frac{(N-1)^{i-1} k_i}{k_1}$, $i=2,3,4$. It should be noted that for

this expansion to be valid, the ϕ_i 's should be bounded for all i , as $N \rightarrow \infty$. This is indeed the case as is easily seen from (2.17). The same is true for the expansions in (3.2) - (3.4) below.

The higher moments of Y are easily obtained from (3.1) by replacing h by $2h$, $3h$, etc. Thus the variance $\sigma^2(h)$ and the third and fourth central moments of Y are obtained in power series in $[(N-1)k_1]^{-1}$ as:

$$\begin{aligned} \sigma^2(h) = & \frac{h^2 \phi_2}{(N-1)k_1} + \frac{h^2(h-1)}{2[(N-1)k_1]^2} [2\phi_3 + (3h-5)\phi_2^2] \\ & + O([(N-1)k_1]^{-3}). \end{aligned} \quad (3.2)$$

$$\mu_3(h) = \frac{h^3}{[(N-1)k_1]^2} [\phi_3 + 3(h-1)\phi_2^2] + O([(N-1)k_1]^{-3}), \quad (3.3)$$

$$\mu_4(h) = \frac{3h^4 \phi_2^2}{[(N-1)k_1]^2} + O([(N-1)k_1]^{-3}). \quad (3.4)$$

If we take

$$h = 1 - k_1 k_3 / 3k_2^2, \quad (3.5)$$

then the leading term in the third moment of Y vanishes and we get an approximately symmetric power transformation of T . Therefore we propose the approximation (\tilde{v} for approximately distributed)

$$\left(\frac{-\log(1-R^2)}{k_1}\right)^h \sim N(\mu_1'(h), \sigma^2(h)), \quad (3.6)$$

where h is given by (3.5) and $\mu_1'(h)$ and $\sigma^2(h)$ are given by (3.1) and (3.2) respectively with the terms shown, i.e., only up to and including the $[(N-1)k_1]^{-2}$ terms.

From (3.6) the approximate probability $\Pr(R \leq x)$ is computed as $\Phi(w)$, where $\Phi(\cdot)$ is the standard normal c.d.f., and

$$w = \left(\left(-\log(1-x^2)/k_1 \right)^h - \mu_1'(h) \right) / \sigma(h) \quad . \quad (3.7)$$

For example, let $N = 50$, $p = 8$, $\rho = .5$, and $x = .7$.

Direct substitution produces

$$k_1 = .4452588, \quad k_2 = .0275862, \quad k_3 = .0025196, \quad \phi_2 = 3.0358143$$

$$\phi_3 = 13.5866127, \quad h = .5085949, \quad \mu_1'(h) = .9821378 \quad .$$

Hence, $w = 1.3164053$, and $\Phi(w) = .9059$; (a computer program for the computation of $\Pr(R \leq x)$ and also the percentiles of R is available from the authors upon request).

4. AN EVALUATION OF THE NORMAL APPROXIMATION

The normal approximation of the previous section was evaluated numerically over a wide range of the parameters p , N and ρ , in terms of both the cumulative probabilities and percentiles. In this section we report the findings regarding the accuracy of the approximation. In general, we conclude the following:

1. The approximation is very accurate for large values of ρ (say $\rho > .5$) with errors in the fourth decimal place of the approximating cumulative probabilities $\Pr(R \leq x)$; if $\rho > .7$, then similar accuracy is obtained for small N even as low as 15.

2. If both, ρ and N are large, then the approximation is extremely accurate especially in the upper tails, with errors $\leq 10^{-4}$ in the upper 1st and 95th percentiles.

3. As is seen from (3.1) and (3.2), the accuracies of the mean $\mu_1'(h)$ and variance $\sigma^2(h)$ of the normal approximation depend on the magnitude of $\theta_N = (N-1)k_1$. The size of θ_N characterizes the accuracy. If θ_N is large, then the approximation of both, proba-

bilities and percentiles is accurate, again with errors in the fourth decimal place.

4. The quality of the approximation of the probabilities $P_r(R \leq x)$ is the same for all x in $(0, 1)$. The approximation is especially simple and it compares in accuracy to Lee's (1971) central and non-central $-F$ approximations.

5. The approximation is not very sensitive to changes in p unless ρ is very small (say .1 or .2).

We now discuss a sample of Tables from the numerical evaluation. Table I gives errors of the approximation for the cumulative probabilities $P_r(R \leq x)$ with sample size N between 16 and 50 and various choices of p and ρ as in Lee (1971). (The first two cases in the Table $N=50$, $p=8$, $\rho=.5$ and $N=50$, $p=6$, $\rho=.5$ are also included in Lee (1971)). The exact probabilities $\Pr(R \leq x)$ as shown were computed using a finite series form of the distribution function of R given in Gurland (1968) for $N-p$ even (see also Johnson and Kotz (1972), ch. 32, Formula (59)). The Table supports the earlier conclusions as can be seen from the errors representing exact minus approximate probabilities times 10^4 . Note the accuracy in the last two cases (ρ large) of the Table. Also, as previously stated the value of θ_N , given at the bottom of the Table, is well indicative of the quality of the approximation. In cases in which $\theta_N > 20$, errors in general occur in the fourth decimal place. In the fourth case, i.e., $N=26$, $p=4$, $\rho=.3$, both ρ and θ_N are small and hence errors appear in the third decimal. Note also that the accuracy of the cumulative probabilities is the same over the whole range of R . As a final comment on Table I we report that in all cases accuracy improves as N increases, while the same accuracy is maintained when p is increased, or decreased, by 2 or 3.

Tables II and III illustrate the performance of the normal approximation in the upper tails in terms of the accuracy of the percentiles x such that $P_r(R \leq x)$ is .99, .95.

Entries under 'Errors' denote 10^4 times the absolute errors of the approximate percentiles as compared with their exact values.

TABLE I
Errors of Approximation for Pr(Rsx) Using Approximation (3.6)

x	N=50, p=.8, ρ=.5		N=50, p=.6, ρ=.5		N=50, p=.6, ρ=.7		N=26, p=4, ρ=.3		N=26, p=4, ρ=.8		N=16, p=4, ρ=.9	
	Exact	Error (*)	Exact	Error	Exact	Error	Exact	Error	Exact	Error	Exact	Error
.1	.0000	0	.0000	0	.0000	0	.0087	-68	.0040	1	.0003	0
.2	.0001	0	.0005	-1	.0000	0	.0680	-38	.0108	3	.0007	1
.3	.0025	-3	.0077	-3	.0000	0	.2157	37	.0282	5	.0018	2
.4	.0293	-1	.0600	5	.0002	0	.4498	43	.0713	5	.0047	5
.5	.1728	9	.2627	10	.0036	0	.7079	-4	.1696	0	.0129	7
.6	.5396	0	.6517	-7	.0460	2	.8988	-16	.3641	-8	.0370	6
.7	.9055	-4	.9446	-1	.3170	0	.9825	-1	.6596	-3	.1108	-6
.8	.9978	0	.9991	0	.8664	0	.9992	0	.9238	3	.3302	-19
.9	1.000	0	1.000	0	.9998	0	1.000	0	.9988	0	.8006	16
0.11	21.81		19.48		38.38		5.69		28.87		28.52	

(*) Entries shown are: (exact - approximate probability) x 10⁴

(**) The values of x for this column are .55 (.05) .95 instead of .1 (.1) .9

TABLE II
 Errors of the 99th Percentiles

N_2	ν	$\rho = .3$		$\rho = .5$		$\rho = .7$		$\rho = .9$	
		Exact	Error ^(*)	Exact	Error	Exact	Error	Exact	Error
20	4	.7573	11	.8326	0	.9047	1	.9700	2
	6	.7862	27	.8489	10	.9125	5	.9721	2
	8	.8088	43	.8623	20	.9192	10	.9740	4
	10	.8270	57	.8735	29	.9249	14	.9756	5
40	4	.6438	5	.7546	4	.8594	1	.9554	1
	6	.6692	0	.7685	1	.8661	0	.9572	1
	8	.6910	6	.7808	1	.8721	1	.9589	1
	10	.7100	12	.7919	4	.8776	2	.9604	1
60	4	.5857	6	.7139	4	.8353	1	.9474	0
	6	.6070	3	.7254	2	.8408	1	.9490	1
	8	.6260	1	.7359	2	.8459	1	.9504	1
	10	.6431	2	.7457	0	.8508	1	.9517	0
100	4	.5235	5	.6696	3	.8086	1	.9385	0
	6	.5394	4	.6781	2	.8127	0	.9396	0
	8	.5541	3	.6861	1	.8165	1	.9407	0
	10	.5678	1	.6936	2	.8202	1	.9417	0
200	4	.4581	2	.6221	1	.7793	0	.9284	0
	6	.4679	2	.6272	1	.7817	1	.9291	0
	8	.4772	2	.6321	1	.7841	0	.9298	0
	10	.4862	2	.6369	1	.7865	0	.9305	0

(*) Entries are 10^4 times the absolute error.

$\nu = p-1$, $N_2 = N-p$.

TABLE III
Errors of the 95th Percentiles

N ₂	v	ρ = .3		ρ = .5		ρ = .7		ρ = .9	
		Exact	Error ^(*)	Exact	Error	Exact	Error	Exact	Error
20	4	.6853	13	.7785	5	.8717	0	.9590	1
	6	.7232	19	.8005	8	.8826	3	.9621	1
	8	.7525	27	.8185	14	.8918	6	.9647	1
	10	.7759	37	.8335	21	.8996	9	.9670	3
40	4	.5783	5	.7044	0	.8282	1	.9448	0
	6	.6089	5	.7215	1	.8365	1	.9471	0
	8	.6350	7	.7367	3	.8441	0	.9493	0
	10	.6574	8	.7502	3	.8510	1	.9512	0
60	4	.5266	2	.6682	1	.8066	0	.9376	0
	6	.5515	2	.6818	0	.8132	0	.9394	0
	8	.5736	4	.6943	0	.8194	0	.9412	0
	10	.5933	4	.7058	1	.8251	0	.9428	0
100	4	.4735	1	.6306	0	.7837	0	.9298	0
	6	.4915	1	.6402	0	.7883	0	.9311	0
	8	.5080	1	.6493	0	.7928	0	.9324	0
	10	.5234	2	.6579	0	.7971	0	.9336	0
200	4	.4198	0	.5918	0	.7397	0	.9215	0
	6	.4306	1	.5974	0	.7624	0	.9222	0
	8	.4408	0	.6028	0	.7650	0	.9230	0
	10	.4505	0	.6081	0	.7676	0	.9237	0

(*) Entries are 10⁴ times the absolute error.

v = p-1, N₂ = N-p

As in Biometrika Tables for Statisticians by Pearson and Hartley (1972), we use labels $N_2 = N-p$ and $v = p-1$. Exact values of the percentiles were kindly supplied to the authors by Professor R. E. Odeh.

From the errors shown in the Tables, it is concluded that the percentiles are very accurate in the upper tail for $\rho > .2$ with a few exceptions when $N_2 < 20$ and $\rho < .5$. As with the cumulative probabilities, large $\rho (\geq .5)$ implies very good accuracy (e.g. the last two columns of the Tables).

5. AN APPLICATION AND CONCLUSIONS

From the earlier literature, e.g. Lee (1971), it may be concluded that the analogue of Fisher's well known z-transformation of the correlation coefficient is inadequate for approximating the distribution of R. It is well known, e.g. see the editorial note accompanying Lee (1971), that this analogue can be improved by appropriate centering. The normal approximation considered in this paper compares favorably with this improvement. It is simpler and compares well with the most accurate among available approximations viz. that based upon the noncentral-F distribution. Moreover, the normal approximation can be readily used for purposes such as construction of confidence intervals for ρ . Other approximations in this context require additional iterative methods. The construction of confidence intervals is now illustrated with an example taken from Pearson and Hartley (1972).

Suppose that in a sample of size $N=50$ from a 5-variate population (i.e. $p = 5$), the observed value of the multiple correlation coefficient is $R = .63$ and that a 90% confidence interval for the correlation ρ is desired. Pearson and Hartley (1972) use this example to illustrate a simple graphical method for the construction of the 90% confidence interval for ρ . The method consists of obtaining $R(.05 | 50, 5, \rho)$ and $R(.95 | 50, 5, \rho)$, the 5th and 95th percentiles of the distribution of R and graphing them against ρ . Two smooth curves are then drawn through them and subsequently the upper and lower confidence limits for ρ are read as the abscissae

TABLE IV
 A Comparison of the Percentiles $R(\alpha|50,5,\rho)$
 and their Approximations $\hat{R}(\alpha|50,5,\rho)$

ρ	.3	.4	.5	.6	.7	.8
$\hat{R}(.05 50,5,\rho)$.2046	.2782	.3702	.4762	.5932	.7194
$R(.05 50,5,\rho)$.2018	.2773	.3703	.4766	.5937	.7202
$\hat{R}(.95 50,5,\rho)$.5626	.6285	.6941	.7588	.8221	.8834
$R(.95 50,5,\rho)$.5614	.6270	.6930	.7581	.8216	.8831

$R(\alpha|50,5,\rho)$ taken from Tables of exact percentiles.
 $\hat{R}(\alpha|50,5,\rho)$ based upon (3.6).

corresponding to the ordinate .63. In Table IV we use the percentiles obtained by linearly interpolating the exact percentiles in Tables provided by Professor R. E. Odeh. These percentiles are accurate to four decimal places as opposed to the ones in *Biometrika* with three decimals. The corresponding lower and upper percentiles obtained through the approximation (3.6) are indicated by $\hat{R}(.05|50,5,\rho)$ and $\hat{R}(.95|50,5,\rho)$ respectively.

The difference between the two sets is clearly miniscule; both sets of percentiles yield the same approximate confidence interval for ρ , (i.e. $.405 \leq \rho \leq .730$) indicating the adequacy of the normal approximation for constructing confidence intervals.

It should be noted that the confidence coefficient is not restricted to be either 90% or 98% and no interpolation is involved in the construction of a confidence interval using the approximation.

ACKNOWLEDGMENTS

We wish to thank the Editor and the Associate Editor for their valuable suggestions for revising the original manuscript. We are also grateful to Professor R. E. Odeh for providing us with tables of the exact percentiles of R.

Research sponsored in part by the Air Force Office of Scientific Research, Air Force Systems Command, USAF under Grant No. AFOSR-77-3360. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

BIBLIOGRAPHY

- Fieller, E. C. (1932). The distribution of the index in a normal bivariate population. Biometrika, 24, 428-440.
- Fisher, R. A. (1928). The general sampling distribution of the multiple correlation coefficient. Proc. Roy. Soc. A, 121, 654-673.
- Gurland, J. (1968). A relatively simple form of the distribution of the multiple correlation coefficient. J.R.S.S., B, 30, 276-283.
- Hodgson, V. (1965). On the sampling distribution of the multiple correlation coefficient (abstract). Ann. Math. Stat. 30, 307.
- Jensen, D. R., Solomon, H. (1972). A Gaussian approximation to the distribution of a definite quadratic form. JASA, 67, 1972.
- Johnson, N. L., Kotz, S. (1972). Continuous Univariate Distributions, 2. John Wiley, New York.
- Khatri, C. G. (1966). A note on a large sample distribution of a transformed multiple correlation coefficient. Ann. Inst. Stat. Math. 18, 375-380.
- Lee, Y. S. (1971). Some results on the sampling distribution of the multiple correlation coefficient. JRSS, B, 117-130.
- Luke, Y. L. (1969). The Special Functions and Their Approximations. Vol. I, Acad. Press. New York.
- Mudholkar, G. S. and Trivedi, M. C. (1981) "A Gaussian Approximation to the Distribution of the Sample Variance for Nonnormal Populations", JASA, 76, No. 374, 479-485.
- Odeh, R. E. (1982). Tables of Percentiles of the Multiple Correlation Coefficient. Private Communication.
- Pearson, E. S., Hartley, H. O. (1972). Biometrika Tables for Statisticians, Vol. 2, Camb. Univ. Press. London.

Watson, G. N. (1918). Asymptotic expansions of hypergeometric functions. Camb. Phil. Soc. Trans. Vol. XXII No. XIV, 277-308.

Wilson, E. B., Hilferty, M. M. (1931). The distribution of chi-square. Proc. Nat. Acad. Sc. 17, 684-8.

Received March, 1982; Revised October, 1982.

Recommended by R. E. Odeh, University of Victoria, Victoria, Canada