SOME NORMALIZING TRANSFORMATIONS OF THE
MULTIPLE CORRELATION COEFFICIENT

by

Vincent Hodgson

FSU Statistics Report M147
ONR Technical Report No. 30

on

THE DEVELOPMENT OF STATISTICAL METHODS
FOR QUALITY CONTROL AND SURVEILLANCE TESTING

December, 1968
Department of Statistics
Florida State University
Tallahassee, Florida

Research supported by the Army, Navy and Air Force under Office of Naval
Research Contract Number NONR-988(08), Task Order NR 042-004 with the
Florida State University Reproduction in whole or in part is permitted
for any purpose of the United States Government.
SOME NORMALIZING TRANSFORMATIONS OF THE MULTIPLE CORRELATION COEFFICIENT

Vincent Hodgson*
The Florida State University, Tallahassee
December, 1968
FSU Statistics Report No. 147

Theorem 1 generalizes Ruben's representation of the correlation coefficient in samples from a bivariate normal population to a representation of the multiple correlation coefficient. Approximation A is a generalization of Ruben's normalizing transformation of the sample correlation coefficient. Two other normalizations of the multiple correlation coefficient, Approximation B and Approximation C, are derived and found by numerical studies to be more accurate. Finally a slightly simpler Approximation C' is recommended.

INTRODUCTION

The starting point for this paper is work by Ruben [10] on the distribution of the sample correlation coefficient in samples of size N from a bivariate normal population. Let \( r \) be the sample correlation coefficient and \( \rho \) be the population correlation coefficient. Let \( \tilde{r} = r/(1-r^2)^{1/2} \) and \( \tilde{\rho} = \rho/(1-\rho^2)^{1/2} \). Ruben shows that \( \tilde{r} \) is distributed as \( (\xi + \tilde{\rho} \chi_{N-1})/\chi_{N-2} \) where \( \xi \) is a standardized normal variate, \( \chi_{N-1} \) and \( \chi_{N-2} \) are chi variates with degrees of freedom equal to their subscripts, and where \( \xi, \chi_{N-1} \) and \( \chi_{N-2} \) are independent. This is an exact result, not an approximation. Using Fisher's approximation to the distribution of the square root of chi square, Ruben finds that

*Dr. Hodgson is currently with the Sperry and Hutchinson Company, New York City. Research for this paper was supported in part by the Office of Naval Research Grant No. NONR 988(08).
\[ \frac{\tilde{r} \left( \tilde{c_i} - \frac{5}{2} \right)^{1/2} - \tilde{r} \left( \tilde{w} - \frac{3}{2} \right)^{1/2}}{(1 + \frac{1}{2} \tilde{r}^2 + p^2)^{1/2}} \]

is approximately distributed as a standardized normal variate and that this approximation is very good.

It is interesting to ask whether Ruben's representation of the sample correlation coefficient can be generalized to a representation of the sample multiple correlation coefficient in samples of size \( N \) from a multivariate normal population and whether this leads to useful approximations to the distribution of the multiple correlation coefficient. The answer to the first question is fairly simple and is given in Theorem 1. The rest of this paper attempts to answer the second question.

Let \( R \) be the sample multiple correlation coefficient and let the Greek capital rho \( \rho \) be the population correlation coefficient and let \( \tilde{R}^2 = R^2 / (1 - R^2) \) and \( \tilde{\rho}^2 = \rho^2 / (1 - \rho^2) \). Let the sample size be \( N \) and \( p \) be the number of variates in the multinormal distribution so that \( R \) is the correlation between one of the normal variates and a linear combination of the \( p - 1 \) other variates.

**Theorem 1.** \( \tilde{R}^2 \) is distributed as \( \chi_{p-2}^2 + \left( \xi + \tilde{\rho} \chi_{N-p}^2 \right)^2 \) where \( \xi \) is a standardized normal variate, \( \chi_{p-2}^2 \), \( \chi_{N-1}^2 \) and \( \chi_{N-p}^2 \) are chi-square variates with degrees of freedom equal to their subscripts, and where \( \xi \), \( \chi_{p-2}^2 \), \( \chi_{N-1}^2 \) and \( \chi_{N-p}^2 \) are independent.

This result can be proved by examining an orthodox derivation of the distribution of \( R \), for example: Anderson [2], page 93. First the distribution of \( (N - p) \tilde{R}^2 / (p - 1) \) given the \( p - 1 \) conditioning variates are fixed is proved to
be noncentral $F$ with $p-1$ and $N-p$ degrees of freedom. Then in the second part of the derivation of the distribution of $R$, the noncentrality parameter is found to be distributed as $\tilde{F}^2$ times a chi-square variate with $N-1$ degrees of freedom.

Gurland [5] has noted and proved the above theorem except that he does not give the explicit representation of the numerator. He proceeds to a very accurate approximation to the distribution of $R$ by assuming that the numeration can be approximated by $g\chi_f^2$ where $g$ and $f$ are constants found by equating the first two moments of $g\chi_f^2$ to those of the numerator. Thus Gurland's approximation is that $\tilde{R}^2$ is distributed as a constant times an $F$ variate with the degrees of freedom of the numerator not, in general, an integer.

This paper considers some normal approximations, suggested directly or indirectly by Theorem 1, to the distribution of the sample multiple correlation coefficient. These approximations are not as accurate, at least for small $N$ and $p$, as Gurland's but are of some theoretical interest and have the advantage of being based on the slightly more familiar and much more thoroughly tabulated normal distribution.

**APPROXIMATION A**

The first approximation suggested by Theorem 1 is a generalization of Ruben's approximation for the bivariate correlation coefficient. For further generality, we consider not merely the square root of chi square but the transformation $(\chi^2)^h$ where, typically but not necessarily, $0<h<1$.

Let the expected value of the numerator in the representation of Theorem 1 be $\mu = (p-1) + (N-1)\tilde{F}^2$. Let $X_1$ be the numerator divided by its expected value:
\[
X_1 = \frac{\chi_{p-2}^2 + (\xi + \bar{\nu} \chi_{N-1})^2}{\mu}.
\]

Similarly let \(X_2 = \frac{\chi_{N-p}^2}{(N-p)}\). By well-known methods (Kendall and Stuart [7], volume 1, page 372) it can be shown that the cumulants of the \(h\)th power of \(X_2\) are:

\[
\kappa_1(X_2^h) = 1 + \frac{h(h-1)}{(N-p)} + \frac{h(h-1)(h-2)(3h-1)}{(N-p)^2} + O((N-p)^{-3})
\]

(1)

\[
\kappa_2(X_2^h) = \frac{2h^2}{(N-p)} + \frac{2h^2(h-1)(3h-1)}{(N-p)^2} + O((N-p)^{-3})
\]

(2)

\[
\kappa_3(X_2^h) = \frac{4h^3(3h-1)}{(N-p)^2} + O((N-p)^{-3})
\]

(3)

\[
\kappa_4(X_2^h) = O((N-p)^{-3}).
\]

(4)

The numerator \(Y_1 = \chi_{p-2}^2 + (\xi + \bar{\nu} \chi_{N-1})^2\) in Theorem 1 is a random variable of the chi-square type in the sense that the \(r\)th cumulant of \(X_1\) is of order \(\mu^{1-r}\). Hence the cumulants of \(X_1^h\) can be calculated in the same way those of \(X_2^h\) have been:

\[
\kappa_1(X_1^h) = 1 + \frac{h(h-1)(2+\bar{\nu}^2)}{\mu} + O(\mu^{-2})
\]

(5)

\[
\kappa_2(X_1^h) = \frac{2h^2(2+\bar{\nu}^2)}{\mu} + O(\mu^{-2})
\]

(6)

\[
\kappa_3(X_1^h) = \frac{4h^3(3h(2+\bar{\nu}^2)^2 - (6+6\bar{\nu}^2+\bar{\nu}^4))}{\mu^2} + O(\mu^{-3})
\]

(7)

\[
\kappa_4(X_1^h) = O(\mu^{-3}).
\]

(8)
When we examine the standardized cumulants $\gamma_1 = \kappa_3 / \kappa_2^{3/2}$ and $\gamma_2 = \kappa_4 / \kappa_2^2$ of $X_1^h$ and $X_2^h$, we find that $\gamma_1(X_1^h)$ is of order $\mu^{-1/2}$ and $\gamma_2(X_1^h)$ is of order $\mu^{-1}$ while $\gamma_1(X_2^h)$ is of order $(N-p)^{-1/2}$ and $\gamma_2(X_2^h)$ is of order $(N-p)^{-1}$. This indicates that $X_1^h$ and $X_2^h$ are approximately normally distributed when $(p-1) + (N-p)\bar{F}^2$ and $(N-p)$, respectively, are large.

Now Geary [4] has shown that $(u_1 - u_2 V) / (\sigma_1^2 + \sigma_2^2) V^{1/2}$ is approximately a standardized normal variate when $V = U_1 / U_2$ is the ratio of two independent normal variates with means $u_1$ and $u_2$ and variances $\sigma_1^2$ and $\sigma_2^2$. Thus from equations (1), (2), (5) and (6), neglecting terms in $(N-p)^{-2}$ in equations (1) and (2), we have a generalization of Ruben's result:

**Approximation A**

$$\frac{\bar{R}^2 h (N-p+h-1)^{h}}{\left[\bar{R}^4 h^2 \sigma_1^2 + \frac{2h^2 (2\bar{F}^2)}{(N-p)^{1-2h}} \right]^{1/2}}$$

is approximately a standardized normal variate.

Choosing $h = 1/2$ simplifies the form of Approximation A and has as its theoretical basis the well-known normal approximation to the distribution of the square root of chi square. Another theoretical consideration is to choose $h$ to reduce the relative skewness of $X_1^h$ or $X_2^h$. The Wilson-Hilferty [11] value of $1/3$ assigned to $h$ will do most to make the distribution of $X_2^h$ symmetric since $\gamma_1(X_2^h) = \{(3h-1)/2(N-p)\} + O((N-p)^{-3/2})$ and this choice of $h$ will have the added advantage of eliminating from the first two cumulants of $X_2^h$ those terms in $(N-p)^{-2}$ which were the leading terms neglected in deriving Approximation A. Similarly equation (7) indicates that to make the distribution of $X_1^h$ most symmetric, one should choose $h = (6+6\bar{F}^2+\bar{F}^4)/3(2+\bar{F}^2)^2$.
which decreases from 1/2 to 1/3 as $\tilde{R}^2$ varies from zero to infinity. However, numerical investigations show that Approximation A is not very sensitive
to the choice of $h$ and that theoretically optimal choices of $h$ do not lead to
the most accurate approximation. The latter point has been noted by Laubacher
[3] who applied the above techniques to the noncentral $F$ and $t$
distributions.

APPROXIMATION B

One major inconvenience of Approximation A is that to use it an awkward
quadratic equation has to be solved. Consider, therefore, a direct
approximation to the distribution of $\tilde{R}^{2h}$. Approximations to the cumulants
of $\tilde{R}^{2h}$ can be extracted from the algebra leading to equations (1)-(8).
Letting $x^h = x^h_1 / x^h_2$, we can find the moments of $x^h - 1$. Ignoring second order
terms we have:

$$E(x^h - 1) = \frac{h(h-1)(2+\tilde{p}^2)}{\mu} + \frac{h(h+1)}{N-p}$$  \hspace{1cm} (9)

$$E(x^h - 1)^2 = \frac{2h^2(2+\tilde{p}^2)}{\mu} + \frac{2h^2}{N-p}.$$ \hspace{1cm} (10)

And ignoring third order terms we have:

$$E(x^h - 1)^3 = \frac{2h^3(9h(2+\tilde{p}^2)^2 - 24 - 24\tilde{p}^2 - 5p^4)}{\mu^2}$$

$$+ \frac{24h^4(2+\tilde{p}^2)}{\mu(N-p)} + \frac{2h^3(9h-5)}{(N-p)^2} \hspace{1cm} (11)$$

$$E(x^h - 1)^4 = \frac{12h^4(2+\tilde{p}^2)^2}{\mu^2} + \frac{24h^4(2+\tilde{p}^2)}{\mu(N-p)} + \frac{12h^4}{(N-p)^2}. \hspace{1cm} (12)$$
Hence for the cumulants of $X^h$ we have, ignoring third order terms:

$$
\kappa_3(X^h) = 4h^3 \left\{ \frac{3h(2+\tilde{p}^2)}{\mu^2} \left( 6+6\tilde{p}^2+\tilde{p}^4 \right) + \frac{3h(2+\tilde{p}^2)}{\mu(N-p)} \right\}
$$

$$
\kappa_4(X^h) = 0.
$$

Equations (10), (13) and (14) indicate a normal approximation to the distribution of $X^h$, and equivalently to that of $\tilde{R}^{2h}$, will be of a quality comparable with normal approximations for the random variables $X_1^h$ and $X_2^h$. Also, for greatest symmetry and perhaps an improved approximation one should put

$$
3h = \frac{(N-p)^2(6+6\tilde{p}^2+\tilde{p}^4) + 4\mu^2}{(N-p)^2(2+\tilde{p}^2)^2 + \mu(N-p)(2+\tilde{p}^2) + \mu^2}.
$$

If $N$ is much larger than $p$ and we consider only terms in $N^2$ in the numerator and denominator then the approximately optimal choice for $h$ is

$$
h^* = \frac{6 + 6\tilde{p}^2 + 5\tilde{p}^4}{3(4+6\tilde{p}^2+3\tilde{p}^4)}
$$

so that $(13+5\sqrt{37})/9(5+\sqrt{37}) = .435 \leq h^* \leq 5/9$ for all possible $\tilde{p}^2$. Clearly considerations of symmetry, not however an infallible guide, indicate choosing $h = 1/2$ and as in Approximation A this choice does most to simplify the new approximation which, from equations (9) and (10), takes the form:

**Approximation B** \(\tilde{R}^{2h}\) is approximately normally distributed with

$$
\text{mean } (\tilde{R}^{2h}) = \left\{ \frac{(p+2h-3) + (N+h-2)\tilde{p}^2}{N - p - h - 1} \right\}^h
$$
\[
\text{variance}(R^{2h}) = \frac{2h^2 (2N-p-1)(1+p^2)}{N-p1_{1+2h}((p-1) + (N-1)p^2)^{1-2h}}
\]

**APPROXIMATION C**

The two approximations above are based on our being able to find and manipulate the moments of the \( h \)th power of random variables related to \( R^2 \) or to functions of \( R^2 \) and the representation of Theorem 1 is used to make the algebra tractable. We now consider a function of \( R^2 \) whose moments can be handled fairly easily because of the form of the probability density function of \( R^2 \). We do not use Theorem 1 except that the consequences of Theorem 1 motivated this part of our investigation. To make the formulas appear shorter and tidier we shall put \( \alpha = \frac{1}{2}(N-1) \) and \( \beta = \frac{1}{2}(p-1) \).

Consider a series expansion of the hypergeometric function form of the probability density function of \( R^2 \):

\[
dG(R^2) = (1-R^2)\alpha \sum_{j=0}^{\infty} \frac{p^2j}{j!} \frac{(R^2)^{\beta-1+j}(1-R^2)^{\alpha-\beta-1}}{\Gamma(\alpha)\Gamma(\alpha-\beta)\Gamma(\beta+j)}
\]

(15)

It can be seen that to take moments of \( R^2 \) leads to a generalized hypergeometric \( _3F_2 \) function. This can be transformed to give Wishart's [12] results but the technique illustrates the natural awkwardness of the moments of \( R^2 \).

If, however, we consider the moments of \( (1-R^2) \) we are led directly to the ordinary hypergeometric \( _2F_1 \) function in a fairly convenient form:

\[
\text{E}(1-R^2)^h = (1-R^2)^{\alpha} \frac{\Gamma(\alpha)\Gamma(\alpha-\beta+h)}{\Gamma(\alpha-\beta)\Gamma(\alpha+h)} _2F_1(\alpha,\alpha;\alpha+h; R^2).
\]

By a well-known transformation formula ([1], equation 15.3.3) for the hypergeometric function we have
\[ E(1-R^2)^h = (1-P^2)^h \frac{\Gamma(\alpha) \Gamma(\alpha-\beta+h)}{\Gamma(\alpha-\beta) \Gamma(\alpha+h)} \binom{\alpha-\beta}{2} (-h, -h, \alpha+h; \frac{p^2}{\alpha}) \]

Using ([1], equation 6.1.47*) we have the approximation

\[ \frac{\Gamma(\alpha) \Gamma(\alpha-\beta+h)}{\Gamma(\alpha-\beta) \Gamma(\alpha+h)} = \frac{(\alpha-\beta)}{\alpha} \left(1 + \binom{h}{2} \frac{\beta}{\alpha(\alpha-\beta)}\right) \]

and we shall take the convenient approximation

\[ \binom{\alpha-\beta}{2} (-h, -h, \alpha+h; \frac{p^2}{\alpha}) = 1 + \frac{h^2 p^2}{\alpha} . \]

Thus we have the expectation formula

\[ E \left[ \frac{\alpha(1-R^2)}{(\alpha-\beta)(1-P^2)} \right]^h = 1 + \binom{h}{2} \frac{\beta}{\alpha(\alpha-\beta)} + \frac{h^2 p^2}{\alpha} . \] (16)

To this order of approximation the third and fourth cumulants of \((1-R^2)^h\) are zero. We therefore suggest:

**Approximation C**

\[ \left[ \frac{(N-1)(1-R^2)}{(N-p)(1-P^2)} \right]^h \]

is approximately normally distributed with

mean

\[ \left[ \frac{(N-1)(1-R^2)}{(N-p)(1-P^2)} \right]^h = \left[ 1 + \frac{(h-1)(p-1)}{(N-1)(N-p)} \right] \]

\[ \left[ \frac{(N-1)(1-R^2)}{(N-p)(1-P^2)} \right]^h \]

variance

\[ \left[ \frac{(N-1)(1-R^2)}{(N-p)(1-P^2)} \right]^h = \frac{2h^2((p-1) + 2(N-p)p^2)}{(N-1)(N-p)} . \]

*The misprint noted by Fama and Roll [3] does not affect our approximation.*
NUMERICAL INVESTIGATIONS

Kraemer [8] has published a table of the 95th percentile of R correct to three decimal places for $P = 0(0.1)0.9$; $p = 7, 9, 11, 13, 17, 21, 25, 31, 35, 41$; and $N = p+10, p+20, p+30$. To each of these 300 percentiles we have obtained 60 approximations by using each of the above three approximations with $\theta = 0.1(0.1)1.0$. We shall illustrate the quality of the normal approximations by a few examples. While one would often use the approximations, in testing hypotheses or constructing confidence intervals, in the upper tail of the distribution of $R$ around the 95th percentile, it would be unwise to rely too heavily on comparisons with Kraemer's results in selecting the best approximation. Approximations which are optimal at the 95th percentile are not necessarily the best in the lower tail or in the center of the distribution.

Table 1 shows for a fairly typical example ($P=5, p=11, N=41$) how the sixty approximations compare with Kraemer's exact result of .784. There are exceptions to most of the useful descriptive statements, including those which follow, one can make about the approximations. Nor does Table 1 illustrate all of the following points.

The approximations are usually most accurate when $h$ is positive and especially when $h$ is between zero and one. Approximation A, besides being the most awkward to work with, tends to be the least accurate and is also the most insensitive to the choice of $h$. However, when these approximations, and most others, are at their worst ($N$ or $\mu = (p-1) + (N-1)p^2$ small), Approximation A is extremely sensitive for values of $h$ between -0.5 and -1.0. Approximations B and C are somewhat alike but Approximation D tends to be more sensitive to changes in $h$ and Approximation C tends to be more accurate.
Table 1. Approximations to the 95\textsuperscript{th} percentile (.784) of R. P=0.5, p=11, N=41.

<table>
<thead>
<tr>
<th>h</th>
<th>Approximation A</th>
<th>Approximation B</th>
<th>Approximation C</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.0</td>
<td>.794</td>
<td>.812</td>
<td>.767</td>
</tr>
<tr>
<td>-.9</td>
<td>.795</td>
<td>.812</td>
<td>.768</td>
</tr>
<tr>
<td>-.8</td>
<td>.795</td>
<td>.811</td>
<td>.768</td>
</tr>
<tr>
<td>-.7</td>
<td>.795</td>
<td>.809</td>
<td>.769</td>
</tr>
<tr>
<td>-.6</td>
<td>.795</td>
<td>.807</td>
<td>.770</td>
</tr>
<tr>
<td>-.5</td>
<td>.795</td>
<td>.805</td>
<td>.771</td>
</tr>
<tr>
<td>-.4</td>
<td>.795</td>
<td>.802</td>
<td>.772</td>
</tr>
<tr>
<td>-.3</td>
<td>.794</td>
<td>.800</td>
<td>.773</td>
</tr>
<tr>
<td>-.2</td>
<td>.794</td>
<td>.798</td>
<td>.774</td>
</tr>
<tr>
<td>-.1</td>
<td>.793</td>
<td>.795</td>
<td>.774</td>
</tr>
<tr>
<td>.1</td>
<td>.792</td>
<td>.791</td>
<td>.776</td>
</tr>
<tr>
<td>.2</td>
<td>.792</td>
<td>.788</td>
<td>.777</td>
</tr>
<tr>
<td>.3</td>
<td>.792</td>
<td>.786</td>
<td>.779</td>
</tr>
<tr>
<td>.4</td>
<td>.792</td>
<td>.784</td>
<td>.780</td>
</tr>
<tr>
<td>.5</td>
<td>.792</td>
<td>.782</td>
<td>.781</td>
</tr>
<tr>
<td>.6</td>
<td>.792</td>
<td>.781</td>
<td>.782</td>
</tr>
<tr>
<td>.7</td>
<td>.792</td>
<td>.779</td>
<td>.783</td>
</tr>
<tr>
<td>.8</td>
<td>.792</td>
<td>.777</td>
<td>.784</td>
</tr>
<tr>
<td>.9</td>
<td>.793</td>
<td>.776</td>
<td>.786</td>
</tr>
<tr>
<td>1.0</td>
<td>.793</td>
<td>.774</td>
<td>.787</td>
</tr>
</tbody>
</table>

A general survey of the results of the computations indicates that a value of h between zero and one would be the best choice and this view is, of course, supported by the theoretical considerations of preceding sections and by the extensive use of statisticians of fractional power normalizations of chi-square, Poisson and other random variables. There is no value of h which is uniformly optimal for all P, p, and N, not even when the criterion of accuracy is based simply on the 95\textsuperscript{th} percentile. Thus it becomes important that the approximation be easy to use. The values of h which make the approximations most convenient to use are h = 1/2 and h=1.

Tables 2a and 2b show the errors of the three approximations when h = 1/2 and when h=1. Each entry in the table is 1,000 \times the difference:
Table 2a. $1000 \times$ Errors in approximations to the $95^{th}$ percentile of $R$, $h = 1/2$ (top row of cells) and $h=1$ (bottom row of cells)

<table>
<thead>
<tr>
<th></th>
<th>p=11, N=21</th>
<th>p=21, N=31</th>
<th>p=31, N=41</th>
<th>p=41, N=51</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>P=0</td>
<td>15 - 8 - 7</td>
<td>8 - 7 - 1</td>
<td>6 - 7 - 1</td>
<td>4 - 6 - 1</td>
</tr>
<tr>
<td></td>
<td>38 -19 7</td>
<td>24 -14 11</td>
<td>18 -11 11</td>
<td>14 - 9 -12</td>
</tr>
<tr>
<td>P=.1</td>
<td>15 - 8 - 7</td>
<td>7 - 8 - 2</td>
<td>5 - 7 0</td>
<td>3 - 6 9</td>
</tr>
<tr>
<td></td>
<td>37 - 19 7</td>
<td>23 -15 10</td>
<td>17 -11 10</td>
<td>13 -10 9</td>
</tr>
<tr>
<td>P=.2</td>
<td>13 - 9 - 7</td>
<td>7 - 8 - 1</td>
<td>5 - 7 0</td>
<td>4 - 6 1</td>
</tr>
<tr>
<td></td>
<td>35 - 19 8</td>
<td>22 -15 10</td>
<td>17 -9 10</td>
<td>13 - 9 9</td>
</tr>
<tr>
<td>P=.3</td>
<td>11 -10 - 5</td>
<td>6 - 8 - 1</td>
<td>4 - 7 0</td>
<td>3 - 6 0</td>
</tr>
<tr>
<td></td>
<td>32 - 19 9</td>
<td>21 -14 11</td>
<td>15 -11 10</td>
<td>12 - 9 8</td>
</tr>
<tr>
<td>P=.4</td>
<td>10 - 9 - 2</td>
<td>6 - 7 1</td>
<td>4 - 6 1</td>
<td>3 - 5 1</td>
</tr>
<tr>
<td></td>
<td>29 -18 12</td>
<td>19 -13 12</td>
<td>14 -10 10</td>
<td>11 - 8 9</td>
</tr>
<tr>
<td>P=.5</td>
<td>7 -10 - 1</td>
<td>4 - 8 0</td>
<td>3 - 6 1</td>
<td>3 - 5 1</td>
</tr>
<tr>
<td></td>
<td>24 -18 12</td>
<td>16 -12 11</td>
<td>12 - 9 9</td>
<td>10 - 7 8</td>
</tr>
<tr>
<td>P=.6</td>
<td>5 - 9 1</td>
<td>4 - 6 1</td>
<td>3 - 5 2</td>
<td>2 - 4 2</td>
</tr>
<tr>
<td></td>
<td>20 -15 13</td>
<td>14 -10 11</td>
<td>11 - 8 9</td>
<td>9 - 6 8</td>
</tr>
<tr>
<td>P=.7</td>
<td>3 - 8 1</td>
<td>2 - 6 1</td>
<td>1 - 5 0</td>
<td>2 - 3 1</td>
</tr>
<tr>
<td></td>
<td>15 -13 8</td>
<td>10 -9 9</td>
<td>7 - 7 7</td>
<td>7 - 5 6</td>
</tr>
<tr>
<td>P=.8</td>
<td>2 - 6 2</td>
<td>2 - 4 1</td>
<td>1 - 3 1</td>
<td>1 - 3 1</td>
</tr>
<tr>
<td></td>
<td>10 -10 10</td>
<td>7 -6 7</td>
<td>6 - 5 6</td>
<td>4 - 4 4</td>
</tr>
<tr>
<td>P=.9</td>
<td>1 - 3 1</td>
<td>1 - 2 1</td>
<td>0 - 2 1</td>
<td>0 - 2 0</td>
</tr>
<tr>
<td></td>
<td>5 - 5 6</td>
<td>4 - 4 4</td>
<td>3 - 3 3</td>
<td>2 - 2 2</td>
</tr>
</tbody>
</table>

approximation to the $95^{th}$ percentile rounded to three decimal places minus Kramer's exact value, also rounded to three decimal places, for the $95^{th}$ percentile. Each cell in the table contains a $2 \times 3$ matrix of six entries, the top row corresponding to $h = 1/2$, the bottom row corresponding to $h=1$, the columns from left to right corresponding to Approximations A, B, and C.

The main purpose of Tables 2a and 2b is to illustrate the accuracy of the six approximations and to indicate that Approximation C with $h = 1/2$
Table 2b. 1000 × Errors in approximations to the 95th percentile of R, h = 1/2 (top row of cells) and h=1 (bottom row of cells)

<table>
<thead>
<tr>
<th></th>
<th>p=11, N=41</th>
<th>p=21, N=51</th>
<th>p=31, N=61</th>
<th>p=41, N=71</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A  B  C</td>
<td>A  B  C</td>
<td>A  B  C</td>
<td>A  B  C</td>
</tr>
<tr>
<td>P=0</td>
<td>27  16 -12</td>
<td>18  8 -6</td>
<td>13 4 -3</td>
<td>9  1 -3</td>
</tr>
<tr>
<td></td>
<td>26  3 -7</td>
<td>20 -1 -1</td>
<td>16 -3 1</td>
<td>12 -4 2</td>
</tr>
<tr>
<td>P=.1</td>
<td>25  14 -12</td>
<td>17  8 -6</td>
<td>12 4 -4</td>
<td>9  1 -2</td>
</tr>
<tr>
<td></td>
<td>24  1 -7</td>
<td>19 -1 -1</td>
<td>15 -3 1</td>
<td>12 -4 2</td>
</tr>
<tr>
<td>P=.2</td>
<td>21  11 -10</td>
<td>15  6 -5</td>
<td>11 3 -3</td>
<td>9  1 -2</td>
</tr>
<tr>
<td></td>
<td>21 -2 -5</td>
<td>17 -3 0</td>
<td>14 -4 2</td>
<td>12 -4 3</td>
</tr>
<tr>
<td>P=.3</td>
<td>16  6 -8</td>
<td>12  3 -5</td>
<td>10 2 -2</td>
<td>8  1 -1</td>
</tr>
<tr>
<td></td>
<td>16 -6 -3</td>
<td>14 -5 1</td>
<td>13 -4 3</td>
<td>11 -4 4</td>
</tr>
<tr>
<td>P=.4</td>
<td>12  2 -6</td>
<td>10  1 -3</td>
<td>7 0 -2</td>
<td>6  1 -1</td>
</tr>
<tr>
<td></td>
<td>12  8 0</td>
<td>12 -6 3</td>
<td>10 -6 3</td>
<td>9 -5 4</td>
</tr>
<tr>
<td>P=.5</td>
<td>8  -2 -3</td>
<td>7  -1 -2</td>
<td>5 2 -1</td>
<td>4  2 0</td>
</tr>
<tr>
<td></td>
<td>9  -10 3</td>
<td>9  -7 4</td>
<td>8 6 4</td>
<td>7  5 4</td>
</tr>
<tr>
<td>P=.6</td>
<td>5  -3 -1</td>
<td>5  -2 0</td>
<td>4 2 0</td>
<td>3  2 0</td>
</tr>
<tr>
<td></td>
<td>7  -10 5</td>
<td>8  -7 5</td>
<td>7 6 5</td>
<td>6  5 4</td>
</tr>
<tr>
<td>P=.7</td>
<td>2  -4 0</td>
<td>2  -3 0</td>
<td>2 3 0</td>
<td>2  2 1</td>
</tr>
<tr>
<td></td>
<td>5  -9 6</td>
<td>5  -7 5</td>
<td>4 6 4</td>
<td>4  4 4</td>
</tr>
<tr>
<td>P=.8</td>
<td>2  -5 2</td>
<td>1  -3 1</td>
<td>1 3 0</td>
<td>1  2 1</td>
</tr>
<tr>
<td></td>
<td>4  -6 6</td>
<td>3  -5 5</td>
<td>3 5 3</td>
<td>2  4 3</td>
</tr>
<tr>
<td>P=.9</td>
<td>0  -3 1</td>
<td>0  -2 1</td>
<td>1 1 1</td>
<td>1  1 1</td>
</tr>
<tr>
<td></td>
<td>1  -4 4</td>
<td>2  -3 3</td>
<td>2 2 3</td>
<td>2  2 3</td>
</tr>
</tbody>
</table>

is the most accurate. Since in most cases the approximate value for the 95th percentile is monotonic in h, one can, from the results tabulated for two values of h, infer how the best value of h varies as the parameter P, p, and N vary; one can also see a little of how sensitive to changes in h the approximations are. Further, when h=1 Approximations B and C are normal approximations to the distributions of $R^2$ and $R^2$ and these can be compared with the other four approximations which are new.
EXAMPLES AND CONCLUDING REMARKS

We suggest that an accurate and convenient approximation to the distribution of the sample multiple correlation coefficient is Approximation C with \( h = 1/2 \) and with the formula for the mean simplified. Thus we are led to:

\[
\text{Approximation } C^* \quad \sqrt{\frac{(N-1)(1-R^2)}{(N-p)(1-P^2)}} \text{ is approximately normally distributed with mean 1 and variance } \frac{(p-1) + 2(N-p)P^2}{2(N-1)(N-p)} .
\]

This Approximation can be used in the testing of hypotheses without confining the statistician to using \( P=0 \) as his null hypothesis. Computing approximate power functions for tests of the multiple correlation coefficient is straightforward with Approximation C’. For example, in testing \( H_0: P=.5 \) against \( H_1: P\neq .5 \) at the .05 level of significance when \( N=100 \) and \( p=6 \), one should accept \( H_0 \) if \( R \) lies between .364 and .654. Table 3 shows some values of the power function computed from Approximation C’ and rounded to three decimal places.

Table 3. Approximate values of the power function of the test \( H_0: P=.5 \) against \( H_1: P\neq .5 \) when \( N=100 \) and \( p=6 \).

<table>
<thead>
<tr>
<th>( P )</th>
<th>0.0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power</td>
<td>1.000</td>
<td>0.980</td>
<td>0.828</td>
<td>0.478</td>
<td>0.163</td>
<td>0.050</td>
<td>0.319</td>
<td>0.886</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Finding confidence intervals for \( P \) requires the solution of an awkward quartic equation. Consider Kramer's example: \( N=19, p=9, R=.875 \). We find that a 90% confidence interval for \( P \) is (.39, .89). By interpolation in his table, Kramer gives (.36,1) as a 95% confidence interval for \( P \). The lower
limits-of-these intervals are comparable. Kramer's tables cannot provide any upper limit other than 1 but we believe that upper limits computed from Approximation C' will be more accurate than lower limits since this approximation, and most others in the literature, are more accurate for larger values of P.

We note that Approximation A is disappointing in view of the accuracy of its precursor developed by Ruben. Approximation B has an analogous precursor: when is large the transformed correlation coefficient is approximately normally distributed in samples of size from a bivariate normal population. The first two moments of are

\[
E(\tilde{r}) = \frac{N-2}{N-3} \tilde{\rho}
\]

\[
\text{Var}(\tilde{r}) = \frac{1}{N-4} + \tilde{\rho}^2 \left\{ \frac{N-1}{N-4} - \frac{(N-2)^2}{(N-3)^2} \right\}
\]

\[
\approx \frac{1}{N-4} + \frac{\tilde{\rho}^2}{N-5}.
\]

Neither Harley [6], who approximated by a noncentral t variate, nor Ruben discusses this normal approximation which seems to be a poor one when applied to the numerical examples Ruben uses.

It seems that Gurland's very accurate approximation is more accurate than Approximation C': much more accurate when and are both small, slightly more accurate in other cases. However, Approximation C' can be used more easily and in more ways since it is based on the well-tabulated normal distribution.
REFERENCES


1. Originating Activity
   The Florida State University
   Department of Statistics.
   Tallahassee, Florida

2a. Report Security Classification
    Unclassified

2b. Group

3. Report Title
   Some Normalizing Transformations of the Multiple Correlation Coefficient

4. Descriptive Notes

5. Author
   Hodgson, Vincent

6. Report Date
   December, 1968

7a. Total No. of pages
   16

7b. No. of References

8a. Contract or Grant No.
    Monr 988(08)

8b. Project No.

8c. Task No.
    NR 042-004

9a. Originator's Report Number(s)
    ONR Technical Report No. 25

9b. Other Report No(s).
    FSU Statistics Report M147

10. Availability/Limitation Notices
    Releasable without limitations on dissemination.

11. Supplementary Notes

12. Sponsoring Military Activity
    Logistics and Mathematical Statistics Branch
    Office of Naval Research
    Washington, D. C. 20360

13. Abstract
   Three normal approximations to the sampling distribution of the multiple correlation coefficient are proposed and investigated. The approximations are not as accurate as Gurland's (A relatively simple form of the distribution of the multiple correlation coefficient, Journal of the Royal Statistical Society, Series B, 30, 276-283) new approximation based on the F distribution but do have the advantage of being based on the more convenient and more thoroughly tabulated normal distribution.

14. Key Words

<table>
<thead>
<tr>
<th>Link A</th>
<th>Link B</th>
<th>Link C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Role Wt.</td>
<td>Role Wt.</td>
<td>Role Wt.</td>
</tr>
<tr>
<td>Multiple correlation coefficient</td>
<td>Normalizing transformations</td>
<td>Distribution theory</td>
</tr>
</tbody>
</table>