International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:5, No:4, 2011

Approximations to the Distribution of the Sample Correlation Coefficient

John N. Haddad and Serge B. Provost

Abstract—Given a bivariate normal sample of correlated variables, (X_i, Y_i), $i = 1, \ldots, n$, an alternative estimator of Pearson's correlation coefficient is obtained in terms of the ranges, $|X_i - Y_i|$. An approximate confidence interval for $\rho_{X,Y}$ is then derived, and a simulation study reveals that the resulting coverage probabilities are in close agreement with the set confidence levels. As well, a new approximant is provided for the density function of R, the sample correlation coefficient. A mixture involving the proposed approximate density of R, denoted by $h_R(r)$, and a density function determined from a known approximation due to R. A. Fisher is shown to accurately approximate the distribution of R. Finally, nearly exact density approximants are obtained on adjusting $h_R(r)$ by a 7^{th} degree polynomial.

Keywords—Sample correlation coefficient, density approximation, confidence intervals.

I. Introduction

CORRELATION between two variables is generally understood to imply a certain departure from stochastic independence. For a discussion on the concept of correlation and certain of its misinterpretations, the reader is referred to [1] and the references therein. The most common measure of correlation between the random variables X and Y is Pearson's (product-moment) correlation coefficient,

$$\rho_{X,Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{\mathrm{E}[(X-\mu_X)^2]\,\mathrm{E}[(Y-\mu_Y)^2]}},$$

where $\mu_X = \mathrm{E}(X)$ and $\mu_Y = \mathrm{E}(Y)$.

Given a random sample $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ from a bivariate normal distribution, $\rho_{X,Y}$ is customarily estimated by the sample correlation coefficient,

$$R = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S_X}\right)\left(\frac{Y_i - \bar{Y}}{S_Y}\right), \qquad (1)$$

where $\bar{X} = \sum_{i=1}^n X_i/n$, $\bar{Y} = \sum_{i=1}^n Y_i/n$, $S_X^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ and $S_Y^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2/(n-1)$.

R. A. Fisher obtained the following representation of the exact density function of R in [2]:

$$f_R(r) = \frac{2^{n-3}}{\pi(n-3)!}(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-2)/4} \times \sum_{i=0}^{\infty} \Gamma^2\Big(\frac{n-i+1}{2}\Big)\frac{(2\,\rho\,r)^i}{i!}, \qquad (2)$$

John Haddad is Associate Professor in the Department of Mathematics and Statistics at Notre Dame University - Louaize, Zouk Mosbeh, Lebanon. E-mail: john.n.haddad@ndu.edu.lb.

Serge Provost is Professor of Statistics in the Department of Statistical & Actuarial Sciences at The University of Western Ontario, London, Canada, N6A 5B7. Corresponding author's e-mail address: provost@stats.uwo.ca.

for $r \in (-1, 1)$. However, this series representation converges very slowly. Fisher's Z-transform is a well known transformation of R whose associated approximate normal distribution possesses some shortcomings, especially when the sample size is small and $|\rho_{X,Y}|$ is large, in which case the distribution of R is markedly skewed. Is was shown in [3] that the normal approximation requires large sample sizes to be valid. Moreover, as mentioned in [4], the variance of R changes with the mean. In the case of bivariate normal vectors, it is known that the asymptotic variance of Fisher's Z statistic does not depend on $\rho_{X,Y}$. However, as was pointed out for instance by [5] and [6], this property does not necessarily carry over to non-normally distributed vectors.

When X and Y follow a bivariate normal distribution with zero means, unit variances and correlation coefficient ρ, and a random sample of size n is available, the statistics being utilized to make inferences about the population correlation coefficient are usually expressed in terms of the products $\{X_iY_i,\ i = 1, \ldots, n\}$. It would appear that, as an alternative, the set of ordered pairs $\{(Z_{1i}, Z_{2i}),\ i = 1, \ldots, n\}$, where Z_{1i} = Min(X_i, Y_i) and Z_{2i} = Max(X_i, Y_i), has yet to be fully exploited for drawing inferences about ρ. It is shown in Section 2 that one can indeed make inferences about ρ from the ranges of the pairs (X_i, Y_i), or equivalently the absolute value of the differences $|X_i - Y_i| = Z_{2i} - Z_{1i} \equiv D_i$. Approximate confidence intervals for ρ, which are based on $|X_i - Y_i|$ and $|X_i + Y_i|$, $i = 1, \ldots, n$, are derived in Section 3. Section 4 proposes two approximations to the density function of R, which turn out to be more accurate than that determined from Fisher's Z statistic.

II. A Range-Based Estimator

Assuming that (X, Y) follows a bivariate normal distribution with zero means, unit variances and correlation coefficient ρ, the joint probability density function of (X, Y) is given by

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\Big\{-\frac{(x^2 - 2\rho xy + y^2)}{2\,(1-\rho^2)}\Big\} \qquad (3)$$

for all real values of x and y. The joint density function of the order statistics (Z_1, Z_2) is then $g_{Z_1,Z_2}(z_1, z_2) = 2!\, f_{X,Y}(z_1, z_2)$ for $-\infty < z_1 < z_2 < \infty$, with $Z_1 = \mathrm{Min}(X, Y)$ and $Z_2 = \mathrm{Max}(X, Y)$, that is,

$$g_{Z_1,Z_2}(z_1,z_2) = \frac{1}{\pi\sqrt{1-\rho^2}} \exp\Big\{-\frac{(z_1^2 - 2\rho z_1 z_2 + z_2^2)}{2\,(1-\rho^2)}\Big\}.$$

The density function of $D = Z_2 - Z_1$ can be obtained as follows. Letting $D = Z_2 - Z_1$ and $Z = Z_1$, one has $Z_2 = D +$

Z and $Z_1 = Z$. The Jacobian of this inverse transformation being 1, the joint density of D and Z is

$$f_{D,Z}(d,z) = \frac{1}{\pi\sqrt{1-\rho^2}} \exp\Big\{ -\frac{2(1-\rho)(z^2+d\,z)+d^2}{2(1-\rho^2)} \Big\},$$

where $-\infty < z < \infty$ and $0 \le d < \infty$.

The marginal density of D is then obtained by integrating out Z as follows:

$$f_D(d) = \int_{-\infty}^{\infty} f_{D,Z}(d,z)dz = \frac{1}{\pi\sqrt{1-\rho^2}} \exp\Big\{ \frac{-d^2}{2(1-\rho^2)} \Big\} \times \int_{-\infty}^{\infty} \exp\Big\{ \frac{-z^2-d\,z}{1+\rho} \Big\} \mathrm{d}z$$

where

$$\begin{aligned} \int_{-\infty}^{\infty} & \exp\Big\{ \frac{-z^2-d\,z}{1+\rho} \Big\} \mathrm{d}z \\ &= \exp\Big\{ \frac{d^2}{4(1+\rho)} \Big\} \int_{-\infty}^{\infty} \exp\Big\{ \frac{-(z+\frac{d}{2})^2}{2(1+\rho)/2} \Big\} \mathrm{d}z \\ &= \exp\Big\{ \frac{d^2}{4(1+\rho)} \Big\} \sqrt{2\,\pi} \sqrt{\frac{1+\rho}{2}}, \end{aligned}$$

since the integrand, $\exp\Big\{-(z+\frac{d}{2})^2/(2(1+\rho)/2)\Big\}$, is proportional to a $\mathcal{N}\left(-d/2\,,\,(1+\rho)/2\right)$ density function, $\mathcal{N}(\mu,\theta)$ denoting a normal distribution with mean μ and variance θ. The density of D is therefore

$$f_D(d) = \frac{1}{\sqrt{\pi}\sqrt{1-\rho}} \exp\Big\{ -\frac{d^2}{4(1-\rho)} \Big\} \tag{4}$$

for $0 \le d < \infty$ and $f_D(d) = 0$ for $d < 0$.

Thus, on the basis of the ranges $d_i = |x_i - y_i|$, $i = 1,\ldots,n$, determined from the pairs of observations $\{(x_1,y_1),\ldots,(x_n,y_n)\}$, the likelihood function is

$$\mathcal{L}(\mathbf{d}\,;\rho) = (\pi(1-\rho))^{-n/2} \exp\Big\{ -\frac{\sum_{i=1}^n d_i^2}{4(1-\rho)} \Big\}, \tag{5}$$

where $\mathbf{d} = (d_1,\ldots,d_n)'$, the loglikelihood function being

$$\ell \equiv \log\mathcal{L} = -\frac{n}{2}\log(\pi(1-\rho)) - \frac{\sum_{i=1}^n d_i^2}{4(1-\rho)}.$$

On setting

$$\frac{d\ell}{d\rho} = \frac{n}{2(1-\rho)} - \frac{\sum_{i=1}^n d_i^2}{4(1-\rho)^2}$$

equal to zero, one has

$$2n(1-\hat\rho) - \sum_{i=1}^n d_i^2 = 0\,,$$

and the m.l.e. of ρ is given by

$$\hat\rho = 1 - \frac{\sum_{i=1}^n d_i^2}{2\,n},$$

that is,

$$\hat\rho = 1 - \frac{\bar d^{\,2}}{2} \tag{6}$$

where

$$\bar d^{\,2} = \frac{\sum_{i=1}^n d_i^2}{n}\,.$$

In order to determine the exact distribution of the corresponding estimator, one may use the fact that $\sum D_i^2 = \sum (X_i - Y_i)^2$ is the sum of the squares of independent $\mathcal{N}(0, 2(1-\rho))$ random variables.

Once a random sample from a bivariate normal distribution whose means and variances are unknown is secured, the variables can be standardized by letting $X_i^* = (X_i - \bar X)/S_X$ and $Y_i^* = (Y_i - \bar Y)/S_Y$. Then, on substituting these standardized variables in Equation (6), one obtains the following representation of the estimator:

$$\begin{aligned} \hat\rho_s &= 1 - \frac{\sum_{i=1}^n (X_i^* - Y_i^*)^2}{2n} \\ &= 1 - \frac{1}{2n}\sum_{i=1}^n \Big(\frac{X_i - \bar X}{S_X} - \frac{Y_i - \bar Y}{S_Y} \Big)^2 \\ &= 1 - \frac{1}{2n}\sum_{i=1}^n \Big(\frac{S_Y\,(X_i - \bar X) - S_X\,(Y_i - \bar Y)}{S_X\,S_Y} \Big)^2 \\ &= 1 - \frac{1}{2n}\Big[\frac{\sum_{i=1}^n (X_i - \bar X)^2}{S_X^2} + \frac{\sum_{i=1}^n (Y_i - \bar Y)^2}{S_Y^2} - \frac{2\sum_{i=1}^n (X_i - \bar X)(Y_i - \bar Y)}{S_X\,S_Y} \Big] \\ &= 1 - \frac{1}{2n}\Big[\frac{(n-1)\,S_X^2}{S_X^2} + \frac{(n-1)\,S_Y^2}{S_Y^2} - 2\,(n-1)\,R \Big] \\ &= 1 - \frac{n-1}{n}\,(1-R) \\ &= R + \frac{1-R}{n}\,. \end{aligned} \tag{7}$$

Consequently, as n increases, $\hat\rho_s$ will tend to R and share its distributional properties. Fisher showed that $\mathrm{E}(R) \simeq \rho - \frac{\rho}{2\,n}\,(1-\rho^2)$. Accordingly, an approximate expression for $\mathrm{E}(\hat\rho_s)$ can be obtained as follows:

$$\begin{aligned} \mathrm{E}(\hat\rho_s) &= \frac{1}{n} + \frac{n-1}{n}\,\mathrm{E}(R) \\ &\simeq \frac{1}{n} + \frac{n-1}{n}\,\rho\left(1 - \frac{1-\rho^2}{2\,n}\right) \\ &= \frac{1}{n} + \frac{(2\,n\,\rho - \rho + \rho^3)(n-1)}{2\,n^2} \\ &= \frac{2\,n^2\,\rho - (3\,n-1)\,\rho + (n-1)\,\rho + 2\,n}{2\,n^2} \\ &= \rho - \Big(\Big(\frac{3\,n-1}{2\,n^2} \Big)\rho - \Big(\frac{n-1}{2\,n^2} \Big)\rho^3 - \frac{1}{n} \Big). \end{aligned}$$

Thus, $\Big(\frac{3\,n-1}{2\,n^2} \Big)\rho - \Big(\frac{n-1}{2\,n^2} \Big)\rho^3 - \frac{1}{n}$ is the approximate bias associated with $\hat\rho_s$.

Note that under the initial distributional assumptions,

$$\frac{|X_i - Y_i|^2}{2\,(1-\rho)} = \frac{(X_i - Y_i)^2}{2\,(1-\rho)} \sim \chi_1^2$$

as $(X_i - Y_i) \sim \mathcal{N}(0, 2(1-\rho))$. This implies that

$$\frac{\bar D^2}{2} = \frac{\sum_{i=1}^n (X_i - Y_i)^2}{2\,n} \sim \frac{1-\rho}{n}\,\chi_n^2\,.$$

Thus,

$$\mathrm{E}(\hat{\rho}) = 1 - \Big(\frac{1-\rho}{n}\Big) n = \rho$$

and

$$\mathrm{Var}(\hat{\rho}) = \Big(\frac{(1-\rho)^2}{n^2}\Big)(2\,n) = \frac{2}{n}\,(1-\rho)^2\,.$$

Observe that the variance will be larger when ρ is negative. This suggests making use of an estimator that is expressed in terms $\sum_{i=1}^{n}(X_i + Y_i^-)^2$ when ρ is negative, where Y_i^- denotes the second component of the i^{th} pair of negatively correlated random variables in a sample of size n. This will result in a variance given in terms of $(1+\rho^-)$ where ρ^- denotes the negative correlation coefficient. Such an estimator can be derived as follows: Let X and Y^- be negatively correlated with correlation coefficient ρ^-, $X \sim \mathcal{N}(0,1)$ and $Y^- \sim \mathcal{N}(0,1)$, and let $Y = -Y^-$ and $\rho = -\rho^-$; then, given a sample $(X_i, Y_i^-)\,,\ i = 1,\ldots,n$, of negatively correlated variables, one can form a sample $(X_i, Y_i)\,,\ i = 1,\ldots,n$, of positively correlated variables and make use of the estimator

$$\hat{\rho} = 1 - \frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{2\,n}\,,$$

which can be re-expressed as

$$1 - \frac{\sum_{i=1}^{n}(X_i + Y_i^-)^2}{2\,n}\,.$$

Then $\widehat{\rho^-}$ is taken to be

$$-\hat{\rho} = \frac{\sum_{i=1}^{n}(X_i + Y_i^-)^2}{2\,n} - 1\,.$$

Since $(X_i + Y_i^-) \sim \mathcal{N}(0, 2(1+\rho^-))$, it follows that $\mathrm{E}(\widehat{\rho^-}) = \rho^-$ and $\mathrm{Var}(\widehat{\rho^-}) = \frac{2}{n}\,(1+\rho^-)^2\,.$

III. Approximate Confidence Intervals for ρ

If one assumes that the random vector (X_i, Y_i) follows a bivariate normal distribution with zero means, equal variances σ^2 and correlation coefficient ρ, then, on noting that $(X_i + Y_i) \sim \mathcal{N}(0, 2\sigma^2(1+\rho))\,,\ (X_i - Y_i) \sim \mathcal{N}(0, 2\sigma^2(1-\rho))$, and that $(X_i + Y_i)$ and $(X_i - Y_i)$ are independently distributed for $i = 1,\ldots,n$, one has

$$\frac{\sum_{i=1}^{n}(X_i + Y_i)^2/(2\sigma^2(1+\rho))}{\sum_{i=1}^{n}(X_i - Y_i)^2/(2\sigma^2(1-\rho))} \sim \mathcal{F}_{n,n}\,. \tag{8}$$

Letting $D_+ = \sum_{i=1}^{n}(X_i + Y_i)^2$ and $D_- = \sum_{i=1}^{n}(X_i - Y_i)^2$, a $100(1-\alpha)\%$ confidence interval for ρ can be determined as follows from the pivotal quantity given in the left-hand side of (8). First, one has

$$\Pr\Big(\mathcal{F}_{1-\frac{\alpha}{2},n,n} < \frac{D_+}{D_-}\,\frac{1-\rho}{1+\rho} < \mathcal{F}_{\frac{\alpha}{2},n,n}\Big) = 1-\alpha$$

or

$$\Pr\Big(\frac{D_-}{D_+}\,\mathcal{F}_{1-\frac{\alpha}{2},n,n} < \frac{1-\rho}{1+\rho} < \frac{D_-}{D_+}\,\mathcal{F}_{\frac{\alpha}{2},n,n}\Big) = 1-\alpha\,.$$

Then, letting $\theta_1 = \frac{D_-}{D_+}\,\mathcal{F}_{1-\frac{\alpha}{2},n,n}$ and $\theta_2 = \frac{D_-}{D_+}\,\mathcal{F}_{\frac{\alpha}{2},n,n}$, where θ_1 and θ_2 are greater than zero and noting that $\theta_1 < \frac{1-\rho}{1+\rho}$ is equivalent to $\rho < \frac{1-\theta_1}{1+\theta_1}$, it follows that

$$\rho < \frac{1 - \frac{D_-}{D_+}\,\mathcal{F}_{1-\frac{\alpha}{2},n,n}}{1 + \frac{D_-}{D_+}\,\mathcal{F}_{1-\frac{\alpha}{2},n,n}}\,.$$

Similarly,

$$\theta_2 > \frac{1-\rho}{1-\rho}$$

leads to

$$\rho > \frac{1 - \frac{D_-}{D_+}\,\mathcal{F}_{\frac{\alpha}{2},n,n}}{1 + \frac{D_-}{D_+}\,\mathcal{F}_{\frac{\alpha}{2},n,n}}\,,$$

so that

$$\Pr\left(\frac{1 - \frac{D_-}{D_+}\,\mathcal{F}_{\frac{\alpha}{2},n,n}}{1 + \frac{D_-}{D_+}\,\mathcal{F}_{\frac{\alpha}{2},n,n}} < \rho < \frac{1 - \frac{D_-}{D_+}\,\mathcal{F}_{1-\frac{\alpha}{2},n,n}}{1 + \frac{D_-}{D_+}\,\mathcal{F}_{1-\frac{\alpha}{2},n,n}}\right) = 1-\alpha\,.$$

Thus,

$$\left(\frac{D_+ - d_-\,\mathcal{F}_{\frac{\alpha}{2},n,n}}{D_+ + d_-\,\mathcal{F}_{\frac{\alpha}{2},n,n}}\,,\ \frac{D_+ - d_-\ \mathcal{F}_{1-\frac{\alpha}{2},n,n}}{D_+ + d_-\ \mathcal{F}_{1-\frac{\alpha}{2},n,n}}\right) \tag{9}$$

is a $100(1-\alpha)\%$ confidence interval for ρ, $\mathcal{F}_{\alpha,n,m}$ denoting the $(100\,(1-\alpha))^{\text{th}}$ percentile of an $\mathcal{F}$ distribution having n and m degrees do freedom.

A simulation study confirmed that the coverage probabilities of this confidence interval are consistently in close agreement with the set confidence levels. Samples of size 50 were generated assuming that $\rho = 0.5$. The coverage probabilities can be readily deduced from the results presented in Table 1.

TABLE I
Number of times $\rho = 0.5$ lies outside the CI's for $n = 50$

No. of CI's	$\alpha = 5\%$	$\alpha = 1\%$
10000	512	112
100000	5048	1037

In practice, it is seldom the case that one will encounter a bivariate data set whose underlying distribution satisfies the assumptions initially made in Section 2. Nevertheless, in terms of the *standardized* variables X_i^* and Y_i^*, one has that

$$\frac{\sum_{i=1}^{n}(X_i^* + Y_i^*)^2/(2(1+\rho))}{\sum_{i=1}^{n}(X_i^* - Y_i^*)^2/(2(1-\rho))} \tag{10}$$

is approximately distributed as an $\mathcal{F}_{n-1,n-1}$ random variable for sufficiently large n. This distributional result can be justified as follows.

Observe that as n gets large, $\mathrm{Var}(X_i^*) \to 1$ and $\mathrm{Var}(Y_i^*) \to 1$. Then, approximately,

$$(X_i^* + Y_i^*) \sim N(0, 2(1+\rho))\,,$$

and

$$(X_i^* - Y_i^*) \sim N(0, 2(1-\rho))\,,$$

and $(X_i^* + Y_i^*)$ and $(X_i^* - Y_i^*)$ are nearly independently

distributed. Thus,

$$\frac{\sum_{i=1}^{n}(X_i^*+Y_i^*)^2}{2\,(1+\rho)} \sim \chi^2_{n-1}\,,\ \text{approximately},$$

one degree of freedom being lost since the mean of $X+Y$ is estimated by $\bar{X}+\bar{Y}$. Similarly,

$$\frac{\sum_{i=1}^{n}(X_i^*-Y_i^*)^2}{2\,(1-\rho)} \sim \chi^2_{n-1}\,,\ \text{approximately}.$$

Accordingly, the ratio given in (10) has approximately an $\mathcal{F}_{n-1,n-1}$ distribution.

It follows from Equation (7) that

$$\begin{aligned}\sum_{i=1}^{n}(X_i^*-Y_i^*)^2 &= \frac{(n-1)\,S_X^2}{S_X^2}+\frac{(n-1)\,S_Y^2}{S_Y^2}-2\,(n-1)\,R\\ &= 2\,(n-1)\,(1-R)\,.\end{aligned}$$

Similarly,

$$\begin{aligned}\sum_{i=1}^{n}(X_i^*+Y_i^*)^2 &= \frac{(n-1)\,S_X^2}{S_X^2}+\frac{(n-1)\,S_Y^2}{S_Y^2}+2\,(n-1)\,R\\ &= 2\,(n-1)\,(1+R)\,.\end{aligned}$$

Thus, one has

$$\begin{aligned}\frac{\sum_{i=1}^{n}(X_i^*+Y_i^*)^2/(2(1+\rho))}{\sum_{i=1}^{n}(X_i^*-Y_i^*)^2/(2(1-\rho))} &= \frac{(1-\rho)}{(1+\rho)}\,\frac{\sum_{i=1}^{n}(X_i^*+Y_i^*)^2}{\sum_{i=1}^{n}(X_i^*-Y_i^*)^2}\\ &= \frac{(1-\rho)}{(1+\rho)}\,\frac{2\,(n-1)\,(1+R)}{2\,(n-1)\,(1-R)}\\ &= \frac{(1-\rho)}{(1+\rho)}\,\frac{(1+R)}{(1-R)}\,,\end{aligned}\qquad(11)$$

which is approximately distributed as an $\mathcal{F}_{n-1,n-1}$ random variable. A derivation analogous to that employed for obtaining the confidence interval given in (9) leads to the following approximate confidence interval for ρ at confidence level $1-\alpha$:

$$\left(\frac{D_+ - D_-\,\mathcal{F}_{\alpha/2,n-1,n-1}}{D_+ + D_-\,\mathcal{F}_{\alpha/2,n-1,n-1}}\,,\ \frac{D_+ - D_-\,\mathcal{F}_{1-\alpha/2,n-1,n-1}}{D_+ + D_-\,\mathcal{F}_{1-\alpha/2,n-1,n-1}}\right),\qquad(12)$$

where $D_+=(1+R)$ and $D_-=(1-R)$.

In a small-scale simulation study, 10,000 and 100,000 samples of size 50 were generated assuming that $\rho=0.5$. The resulting coverage probabilities can be deduced from the results included in Table 2.

TABLE II
NUMBER OF TIMES $\rho=0.5$ LIES OUTSIDE THE CI'S FOR $n=50$

No. of CI's	$\alpha=5\%$	$\alpha=1\%$
10000	464	79
100000	4810	915

IV. Alternative Density Approximations for R

Two approximations to density function of R are proposed in this section. The first one is obtained by applying the change of variable technique to the quantity specified by Equation (11). Let $u(\rho)=\frac{1-\rho}{1+\rho}$, and $x=u(\rho)\,\frac{1+r}{1-r}$, which, which, as explained in the previous section, is approximately distributed as an $\mathcal{F}_{n-1,n-1}$ random variables. Since the probability density function of the $\mathcal{F}_{n,m}$ distribution is

$$f_{n,m}(x)=\frac{\Gamma\Big(\frac{n+m}{2}\Big)\Big(\frac{n}{m}\Big)^{\frac{n}{2}}\,x^{\frac{n}{2}-1}}{\Gamma\Big(\frac{n}{2}\Big)\Gamma\Big(\frac{m}{2}\Big)\Big(1+\frac{n\,x}{m}\Big)^{\frac{n+m}{2}}}\,,$$

one has

$$f_{n-1,n-1}(x)=\frac{\Gamma(n-1)\,x^{\frac{n-3}{2}}}{\Big(\Gamma\Big(\frac{n-1}{2}\Big)\Big)^2\,(1+x)^{n-1}}\,.$$

Noting that $\frac{dx}{dr}=\frac{2\,u(\rho)}{(1-r)^2}$, the resulting approximation to the density function of R is given by

$$h_R(r)=\frac{2\,\Gamma(n-1)}{\Big(\Gamma\Big(\frac{n-1}{2}\Big)\Big)^2}\,u(\rho)^{\frac{n-1}{2}}\,\frac{(1-r^2)^{\frac{n-3}{2}}}{(1-r+u(\rho)\,(1+r))^{n-1}}\,.\qquad(13)$$

Alternatively, an approximate density function can be derived as follows from Fisher's Z-transform, that is, $Z=\frac{1}{2}\ln\Big(\frac{1+R}{1-R}\Big)$. Let

$$z=\frac{1}{2}\ln\Big(\frac{1+r}{1-r}\Big),$$

so that

$$r=\frac{e^{2z}-1}{e^{2z}+1}.$$

Then,

$$\frac{dr}{dz}=\frac{2e^{2z}}{e^{2z}+1}-\frac{2e^{2z}(e^{2z}-1)}{(e^{2z}+1)^2}=\frac{4e^{2z}}{(e^{2z}+1)^2}\,,$$

$$(1-r^2)^{\frac{n-3}{2}}=\Big[1-\Big(\frac{e^{2z}-1}{e^{2z}+1}\Big)^2\Big]^{\frac{n-3}{2}}=\Big(\frac{4e^{2z}}{(e^{2z}+1)^2}\Big)^{\frac{n-3}{2}},$$

$$(1-r+u(\rho)(1+r))^{n-1}=\Big(1-\frac{e^{2z}-1}{e^{2z}+1}+u(\rho)\Big(1+\frac{e^{2z}-1}{e^{2z}+1}\Big)\Big)^{n-1}$$

and the density of Z is

$$g_Z(z)\ =\ \frac{2\,\Gamma(n-1)}{(\Gamma(\frac{n-1}{2}))^2}\Big(\frac{e^z(u(\rho))^{\frac{1}{2}}}{1+u(\rho)e^{2z}}\Big)^{n-1}.$$

Since $u(\rho)=\frac{1-\rho}{1+\rho}$, the density of Z can also be expressed as follows:

$$\begin{aligned}g_Z(z)&=\frac{2\,\Gamma(n-1)}{(\Gamma(\frac{n-1}{2}))^2}(u(\rho))^{\frac{n-1}{2}}\,\frac{(\frac{4e^{2z}}{(e^{2z}+1)^2})^{\frac{n-3}{2}}}{(\frac{2(1+u(\rho)e^{2z})}{e^{2z}+1})^{n-1}}\,\frac{4e^{2z}}{(e^{2z}+1)^2}\\ &=\frac{2\,\Gamma(n-1)}{(\Gamma(\frac{n-1}{2}))^2}\Big(\frac{e^z}{1+u(\rho)e^{2z}}\Big)^{n-1}(u(\rho))^{\frac{n-1}{2}}\\ &=\frac{2\,\Gamma(n-1)}{(\Gamma(\frac{n-1}{2}))^2}\Big(\exp\Big(-z+\frac{1}{2}\ln\Big(\frac{1+\rho}{1-\rho}\Big)\Big)\\ &\qquad+\exp\Big(z-\frac{1}{2}\ln\Big(\frac{1+\rho}{1-\rho}\Big)\Big)\Big)^{-n+1}.\end{aligned}\qquad(14)$$

Clearly, as defined above, Z is not normally distributed, which is consistent with a remark made by [7]. Nevertheless,

we observed that Z tends to a normal distribution with mean $\frac{1}{2}\{\ln[(1+\rho)/(1-\rho)]\}$ and variance $1/(n-2)$. For comparison purposes, $Z^* = \frac{1}{2}\{\ln[(1+R)/(1-R)]\}$, that is, Fisher's Z-transform applied to R, is known to be asymptotically distributed as a $\mathcal{N}(\frac{1}{2}\{\ln[(1+\rho)/(1-\rho)]\}, 1/(\mathrm{n}-3))$ random variable. Upon inversion via the change of variable technique, an approximation to the density function of $R = (e^{2Z^*} - 1)/(e^{2Z^*}+1)$ can be obtained as follows. Since

$$z^* = \frac{1}{2}\ln\Big(\frac{1+r}{1-r}\Big),$$
$$\frac{dz^*}{dr} = \frac{1}{1-r^2},$$

and given that the approximate density of Z^* is

$$f(z^*) = \frac{\sqrt{n-3}}{\sqrt{2\pi}}\exp\Big(-\frac{n-3}{2}\Big(z^* - \frac{1}{2}\ln\Big(\frac{1+\rho}{1-\rho}\Big)\Big)^2\Big),$$

one has the following approximate density function for R:

$$g^*(r) = \frac{\sqrt{n-3}}{\sqrt{2\pi}(1-r^2)}\exp\Big(-\frac{n-3}{2}\Big(\frac{1}{2}\ln\Big(\frac{1+r}{1-r}\Big) - \frac{1}{2}\ln\Big(\frac{1+\rho}{1-\rho}\Big)\Big)^2\Big). \quad (15)$$

Interestingly, an equal mixture of the approximate densities given in (13) and (15) provides more accurate approximations than either one of them, as (15) overestimates the variance while (13) underestimates it. This is graphically illustrated in Figures 1 and 2.

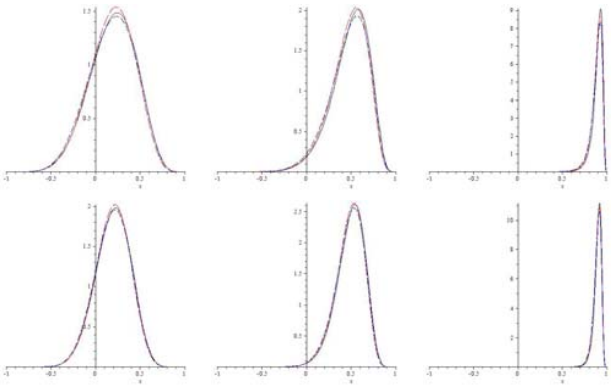

Fig. 1. Exact density of R from Equation (2): solid line, and two approximate densities from Equation (13): long dashes and Equation (15): dashed line, for $\rho = 0.2,\ 0.5$ and 0.9 (left to right) and sample sizes 15 and 25 (top and bottom graphs).

Another approximation to the density of R is now obtained by multiplying the proposed approximate density $h_R(r)$ by $p_d(r)$, a polynomial of degree d, so that the first d moments of the resulting density,

$$hp_d(r) = h_R(r)\, p_d(r), \quad (16)$$

coincide with those of R. This approach is discussed for instance in [8]. Letting $p_d(r) = \sum_{j=0}^{d} \xi_j\, r^j$, the coefficients ξ_j are determined as follows, assuming a polynomial adjustment of degree $d = 7$.

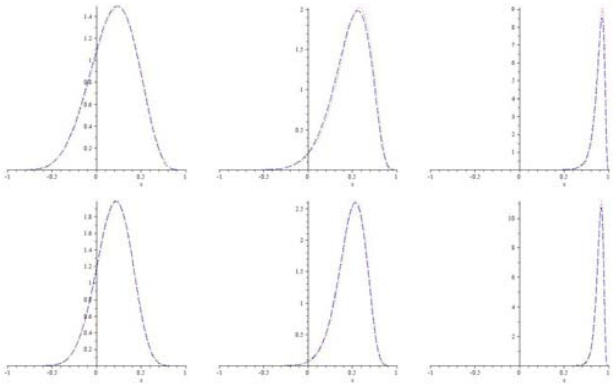

Fig. 2. Exact density of R from Equation (2): dotted line, and mixture of the approximate density functions specified by (13) and (15): dashed line for $\rho = 0.2,\ 0.5$ and 0.9 (left to right) and sample sizes 15 and 25 (top and bottom graphs).

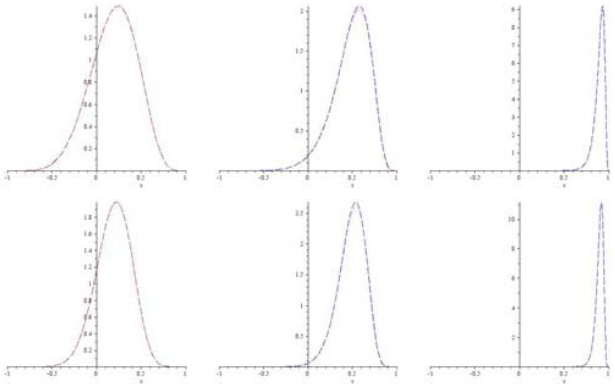

Fig. 3. Exact density of R from Equation (2): dotted line, and the approximate density from Equation (16) with $d = 7$: dashed line) for $\rho = 0.2,\ 0.5$ and 0.9 (left to right) and sample sizes 15 and 25 (top and bottom graphs).

First, letting m_i denote the i^{th} moment of the distribution specified by $h_R(r)$, we evaluate the 8×8 matrix $\mathcal{M}$ whose j^{th} row is $(m_{j-1}, m_j, \ldots, m_{j+6})$, $j = 1, 2, \ldots, 8$, as well as its inverse $\mathcal{M}^{-1}$. We then multiply $\mathcal{M}^{-1}$ by $(\mu_0, ..., \mu_7)'$, the vector of exact moments of R, in order to determine the polynomial coefficients $(\xi_0, ..., \xi_7)'$. The resulting approximate density, that is, $hp_7(r) = h_R(r)\sum_{j=0}^{7}\xi_j r^j$, is plotted in Figure 3 for certain values of ρ and n. Manifestly, this approximation proves to be remarkably accurate.

ACKNOWLEDGMENT

The second author wishes to acknowledge the financial support of the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] A. M. Mathai, The concept of correlation and misinterpretations. *International Journal of Mathematical and Statistical Sciences*, 1998, **7**: 157–167.
[2] R. A. Fisher, Distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 1915, **10**: 507–521.
[3] A. Winterbottom, A note on the derivation of Fisher's transformation of the correlation coefficient. *The American Statistician*, 1979, **33**: 142–143.
[4] H. Hotelling, New light on the correlation coefficient and its transforms. *Journal of Royal Statistical Society, Ser. B*., 1953, **15**: 193–232.
[5] A. K. Gayen, The frequency distribution of the product-moment correlation coefficient in random samples of any size drawn from non-normal universes. *Biometrika*, 1951, **38**: 219–247.
[6] D. L. Hawkins, Using U statistics to derive the asymptotic distribution of Fisher's Z statistic. *The American Statistician*, 1989, **43**: 235–237.
[7] S. Konishi, An approximation to the distribution of the sample correlation coefficient. *Biometrika*, 1978, **65**: 654–656.
[8] H.-T. Ha and S. B. Provost, A viable alternative to resorting to statistical tables. *Communications in Statistics–Simulation and Computation*, 2007, **36**: 1135–1151.