Statistica Sinica 1(1991), 493-502

ON THE DISTRIBUTION OF THE MULTIPLE CORRELATION COEFFICIENT AND THE KINKED CHI-SQUARE RANDOM VARIABLE

John Gurland and Osebekwin Asiribo

University of Wisconsin

Abstract: **In a previous article (Gurland (1968)), a relatively simple form of the distribution of the multiple correlation coefficient R was presented in the form of a mixture of scaled Beta distributions; and an approximation of the distribution was suggested in the form of a scaled F distribution. In a subsequent article (Gurland and Milton (1970)) other representations of the distribution of R were presented and investigated. In the present article the role of the "kinked χ^2" distribution in this problem is emphasized, to yield other possible representations of the distribution of R.**

Key words and phrases: **Multiple correlation, series representations of distributions, modified chi-square distribution.**

1. Introduction

Let X be a p-dimensional random vector with non-singular covariance matrix $\Sigma = [\sigma_{ij}]$. With no loss of generality we consider the multiple correlation coefficient

$$\overline{R}_{1.23\ldots p} = \left(\frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{\sigma_{11}}\right)^{1/2} = \overline{R},$$

say, where

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}; \qquad \begin{aligned} \Sigma_{11} &= \sigma_{11} \\ \Sigma_{12} &= [\sigma_{12}\sigma_{13}\ldots\sigma_{1p}]. \end{aligned}$$

Let $[X_{i\alpha}]$ be a $p \times N$ random sample from the above distribution and

$$A = [a_{ij}] = \left[\sum_{\alpha=1}^{N}(X_{i\alpha} - \overline{X}_i)(X_{j\alpha} - \overline{X}_j)\right] = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

where $\overline{X}_i = \frac{1}{N}\sum_{\alpha=1}^{N} X_{i\alpha}$, $A_{11} = a_{11}$ and $A_{12} = [a_{12}a_{13}\ldots a_{1p}]$. The sample multiple correlation coefficient is defined by

$$R_{1.23\ldots p} = \left(\frac{A_{12}A_{22}^{-1}A_{21}}{a_{11}}\right)^{1/2} = R, \quad \text{say.}$$

It has been shown (Gurland (1968)) that if X has a p-variate normal distribution, then $U = R^2/(1-R^2)$ is distributed as a ratio Y_1/Y_2 of independent random variables with characteristic functions

$$\phi_{Y_1}(t) = \frac{(1-2it)^k}{(1-2iat)^h}; \quad \phi_{Y_2}(t) = (1-2it)^{-k}$$

where

$$k = \frac{N-p}{2}; \quad h = \frac{N-1}{2}; \quad \theta = \frac{\overline{R}^2}{1-\overline{R}^2}; \quad a = 1+\theta = \frac{1}{1-\overline{R}^2}.$$

Hodgson (1968) has noted that U is distributed as

$$\frac{(\sqrt{\theta}\chi_{N-1} + Z)^2 + \chi_{p-2}^2}{\chi_{2k}^2} \tag{1}$$

where χ, χ^2 denote chi, chi-square random variables, respectively, with degrees of freedom as indicated, and Z is a standard normal random variable. All the random variables appearing in (1) are independent. It can be shown (Lee (1971)) that the numerator in (1) is distributed as the random variable Y_1 above with characteristic function as indicated. A particular case of (1), with $p = 2$, given by

$$\frac{r}{\sqrt{1-r^2}} \sim \frac{\frac{\rho}{\sqrt{1-\rho^2}}\chi_{N-1} + Z}{\chi_{N-2}}$$

was presented by Elfving (1947) and Ruben (1966). Here r and ρ are the sample and population correlation coefficients, respectively, in a bivariate distribution.

From the characteristic function ϕ_{Y_1} or from the numerator of (1) it is clear that Y_1 reduces to a χ_{p-1}^2 random variable when $\theta = 0$ ($\overline{R} = 0$). Y_1 is an interesting random variable per se; and for convenience we shall refer to it here as a "kinked χ^2", in which the kink may be regarded as $\theta\chi_{N-1}$. When this kink vanishes ($\theta = 0$) the distribution of Y_1 is that of χ_{p-1}^2.

The purpose of this paper is to present various expressions for the distribution of Y_1, and hence for that of R.

2. Various Representations of the Distribution of Y_1 (kinked χ^2)

Some forms of the distribution developed by Gurland (1968) and Gurland and Milton (1970) are recapitulated here for convenience in the following Sections 2.1 and 2.2.

2.1. Finite series representation for the case k a positive integer

As in Gurland (1968) we can write

$$\phi_{Y_1}(t) = \left(\frac{1}{1-2iat}\right)^{\frac{p-1}{2}} \left(\frac{1}{a}\right)^k \sum_{j=0}^{k} \binom{k}{j} \left(\frac{\theta}{1-2iat}\right)^j.$$

Hence, the distribution functions F_{Y_1}, F_U are given by

$$F_{Y_1}(x) = \sum_{j=0}^{k} b_j F_{p-1+2j}(x/a) \tag{2}$$

$$F_U(x) = \sum_{j=0}^{k} b_j F_{p-1+2j,2k}(x/a) \tag{3}$$

where b_j is the binomial coefficient

$$b_j = \binom{k}{j} \left(\frac{a-1}{a}\right)^j (1/a)^{k-j}. \tag{4}$$

Here F_v denotes the distribution function of χ_v^2 and F_{v_1,v_2} that of a ratio of independent random variables $\chi_{v_1}^2/\chi_{v_2}^2$. The distribution function F_R is readily obtained from the relation

$$F_R(x) = F_{R^2}(x^2) = F_U\left(\frac{x^2}{1-x^2}\right). \tag{5}$$

For the case k a positive integer, Fisher (1928) has also developed a finite series for the distribution of R. Although the series is finite, each term in it is a hypergeometric function. For computational purposes the series (3) above, in conjunction with the relations (5), is much simpler.

2.2. General family of series expansions for $F_{Y_1}(x)$ in scaled χ^2 distribution functions

According to Gurland and Milton (1970), the characteristic function of the kinked χ^2 random variable can be expressed as

$$\phi_{Y_1}(t) = \frac{(bz)^{\frac{p-1}{2}}}{a^{\frac{N-1}{2}}} \sum_{j=0}^{\infty} c_j z^j \tag{6}$$

where

$$z = \frac{1}{1-2ibt}; \quad c_j = \begin{cases} \sum_{r=0}^{j} v_r \delta_{j-r}, & N-p \text{ odd} \\ \sum_{r=0}^{\min(j,k)} v_r \delta_{j-r}, & N-p \text{ even} \end{cases} \quad j = 0,1,2,\dots ;$$

$$v_j = \binom{k}{j}(b-1)^j; \qquad \delta_j = \binom{-\frac{N-1}{2}}{j}\left(\frac{b-a}{a}\right)^j .$$

The series in (6) converges for b satisfying $0 < b \leq 2$ and $0 < b/a \leq 2$. Term by term inversion of this series yields a mixture of scaled χ^2 distributions for the distribution function of Y_1:

$$F_{Y_1}(x) = \frac{b^{\frac{p-1}{2}}}{a^{\frac{N-1}{2}}} \sum_{j=0}^{\infty} c_j F_{p-1+2j}(x/b).$$

As in Section 2.1, the distribution functions of U and R follow readily. The effect of various choices of b, in the above series, on computing the distribution function of R has been examined by Gurland and Milton (1970). For k an integer the optimal choice is $b = a$; but for k fractional the optimal choice appears to be $b = a$ when $\bar{R}^2 \leq 1/2$ but $b = 2$ when $\bar{R}^2 > 1/2$. The case $b = 1$ is also interesting in that the coefficients in the series are probabilities of a negative binomial distribution, but this series converges more slowly than for the above choices indicated.

2.3. Finite series of confluent hypergeometric functions for $f_{Y_1}(x)$ when k fractional (N odd, p even) or (N even, p odd)

The characteristic function of the kinked chi-square random variable Y_1 can be written as

$$\begin{aligned} \phi_{Y_1}(t) &= \left(\frac{1-2it}{1-2iat}\right)^{k+\frac{1}{2}} \left(\frac{1}{1-2iat}\right)^{\frac{p-2}{2}} \left(\frac{1}{1-2it}\right)^{1/2} \\ &= \sum_{j=0}^{k+\frac{1}{2}} b'_j \left(\frac{1}{1-2iat}\right)^{j+\frac{p-2}{2}} \left(\frac{1}{1-2it}\right)^{1/2} \end{aligned} \tag{7}$$

where b'_j is the same as b_j in (4) but with k replaced by $k + 1/2$. Let $W_j = a\chi^2_{p-2+2j} + \chi^2_1$, a weighted sum of independent chi-square random variables as indicated. Then (7) can be written as

$$\phi_{Y_1}(t) = \sum_{j=0}^{k+\frac{1}{2}} b'_j \phi_{W_j}(t) \tag{8}$$

where ϕ_{W_j} is the characteristic function of W_j. From Erdélyi (1954)

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}(\beta - it)^{-\mu}(\gamma - it)^{-v}e^{-itx}dt = \frac{e^{-\beta x}x^{\mu+v-1}}{\Gamma(\mu+v)}{}_1F_1\{v, v+\mu, (\beta-\gamma)x\},$$

$$x > 0, \quad \operatorname{Re}\gamma > 0, \quad \operatorname{Re}(\mu+1) > v,$$

where ${}_1F_1$ is the confluent hypergeometric function. This function has series representation

$${}_1F_1(b,c,z) = \sum_{j=0}^{\infty}\frac{(b)_j}{(c)_j}\frac{z^j}{j!}, \qquad \begin{array}{l} c \neq -m, \quad (b)_j = b(b+1)\cdots(b+j-1), \\ m = 0, 1, 2, \dots, \end{array}$$

convergent for all z. Consequently the p.d.f. (probability density function) of W_j is given by

$$\begin{aligned} f_{W_j}(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\phi_{W_j}(t)e^{-itx}dt \\ &= \left(\frac{1}{2a}\right)^{j+\frac{p-2}{2}}\frac{e^{-x/(2a)}x^{j+(p-1)/2}}{\sqrt{2}\Gamma\left(j+\frac{p-1}{2}\right)}{}_1F_1\left\{\frac{1}{2}, j+\frac{p-1}{2}, -\frac{a-1}{2a}x\right\}. \end{aligned}$$

Thus, by inversion of (8), the p.d.f. of the kinked χ^2 random variable Y_1 can be expressed as

$$f_{Y_1}(x) = \sum_{j=0}^{k+\frac{1}{2}} d_j e^{-x/(2a)}x^{j+(p-1)/2}{}_1F_1\left\{\frac{1}{2}, j+\frac{p-1}{2}, -\frac{a-1}{2a}x\right\}$$

where $d_j = b_j'(2a)^{-(j+\frac{p-2}{2})}\left[\sqrt{2}\Gamma\left(j+\frac{p-1}{2}\right)\right]^{-1}$.

2.4. Expression for $f_{Y_1}(x)$ in terms of a single confluent hypergeometric function

The Laplace transform of Y_1 can be written as

$$\int_0^{\infty} e^{-px}f_{Y_1}(x)dx = \frac{(1+2p)^k}{(1+2ap)^h}.$$

From Erdélyi (1954) the inversion of the Laplace transform

$$\int_0^{\infty} f(x)e^{-px}dx = \Gamma(2v-2\lambda)(p-\alpha)^{2\lambda}(p-\beta)^{-2v}, \quad \operatorname{Re}(v-\lambda) > 0,$$

yields

$$f(x) = x^{2v-2\lambda-1}e^{\alpha x}{}_1F_1\{2v, 2v-2\lambda, (\beta-\alpha)x\}, \quad x > 0.$$

Thus

$$f_{Y_1}(x) = \frac{2^{k-h}a^{-h}}{\Gamma(h-k)} x^{h-k-1} e^{-x/2} {}_1F_1\Big\{h, h-k, \frac{a-1}{2a}x\Big\}, \quad x > 0.$$

Application of Kummer's relation ${}_1F_1(c,d,z) = e^z {}_1F_1(d-c,d,-z)$ and further simplification results in the representation

$$f_{Y_1}(x) = \frac{2^{-\frac{p-1}{2}}a^{-h}}{\Gamma(\frac{p-1}{2})} x^{(p-3)/2} e^{-x/(2a)} {}_1F_1\Big\{-k, h-k, -\frac{a-1}{2a}x\Big\}, \quad x > 0.$$

For k a positive integer this yields a finite series of scaled χ^2 distributions, and for k fractional, an infinite such series. This leads accordingly to a finite or infinite series of scaled Beta distributions for the distribution of R.

3. An Approximation to the Distribution of R Based on Moments of Y_1

In Gurland (1968) the approximation $Y_1 \approx g\chi_f^2$ was suggested, by equating the first two moments of both sides. Thus

$$g = \frac{ha^2 - k}{ha - k}; \qquad f = \frac{2(ha-k)^2}{ha^2 - k}. \tag{9}$$

Then the approximation

$$U \approx \frac{g\chi_f^2}{\chi_{2k}^2} \tag{10}$$

is applied to compute approximate values of $F_R(x)$ by utilizing the relation in (5). This approximation apparently works rather well as evident from the numerical investigation by Gurland and Milton (1970).

A further approximation can be obtained by applying the Wilson-Hilferty transformation of a χ^2 random variable as in Kendall and Stuart (1969). From (10) write

$$\frac{2kU}{fg} = \frac{\chi_f^2}{\chi_{2k}^2}\frac{2k}{f} = F_{f,2k}.$$

Then

$$t(F) = \frac{\left(1 - \frac{1}{9k}\right)F_{f,2k}^{1/3} - \left(1 - \frac{2}{9f}\right)}{\sqrt{\frac{1}{9k}F_{f,2k}^{2/3} + \frac{2}{9f}}} \tag{11}$$

is distributed approximately as a standard normal random variable. Thus

$$F_U(x) \approx \phi\Big\{t\Big(\frac{2kx}{fg}\Big)\Big\} \tag{12}$$

where ϕ is the distribution function of a standard normal random variable. Then $F_R(x)$ is obtained through relations (5).

This approximation is attractive from a practical standpoint as it is based on the normal distribution. Its behavior is similar to that of the approximation based on (10), as can be seen from the illustrative Tables 1 and 2. These tables correspond to $p = 6, 10$ respectively, with $N = 10, 20, 40$ in Table 1 and $N = 14, 20, 40$ in Table 2. The grid of parameter values is $\overline{R} = 0$, .1(.2).9 and of x values is $x = .1(.2).9$. The exact values of $F_R(x)$ are given, along with errors ($\times 10^4$) of approximate values. For comparison, the errors based on using approximation (10) are also included. It is evident from these tables that sometimes the absolute errors based on approximation (12) are greater, sometimes less, but for the most part both approximations are quite similar, as expected. The approximation based on (12), however, has the advantage of being obtainable from tables of the standard normal distribution.

4. Conclusion

Although the distribution of the multiple correlation coefficient is expressible in finite series for integral values of k and in convergent infinite series for k fractional, simple approximations are desirable from a practical standpoint. It is seen that the kinked χ^2 distribution plays a key role in the distribution of R, and in the development of approximations for the distribution of U and hence of R. This leads to some suggested approximations, some of which are based on the normal distribution.

Acknowledgments

This research was supported in part by the Wisconsin Alumni Research Foundation, Project # 140410. We are also grateful for preliminary computations by Michael Wincek in examining some of the suggested approximations.

Table 1. Exact values of $F_R(x)$ and errors ($\times 10^4$) of approximate values for $p = 6$ and $N = 10, 20, 40$ (Error = exact − approximate)

	$\bar{R}$ 0.0			0.1			0.3			0.5			0.7			0.9		
		Error using	Error using		Error using	Error using		Error using	Error using		Error using	Error using		Error using	Error using		Error using	Error using
N	Prob.	(12)	(10)	Prob.	(12)	(10)	Prob.	(12)	(10)	Prob.	(12)	(10)	Prob.	(12)	(10)	Prob.	(12)	(10)
10	0.0000	−1	0	0.0000	−1	0	0.0000	−1	0	0.0000	0	0	0.0000	0	0	0.0000	0	0
20	0.0005	−6	0	0.0004	−5	1	0.0002	−2	1	0.0000	0	0	0.0000	0	0	0.0000	0	0
40	0.0036	−15	0	0.0030	−13	1	0.0006	0	4	0.0000	0	0	0.0000	0	0	0.0000	0	0
10	0.0080	−13	0	0.0076	−13	1	0.0055	−10	2	0.0025	−4	4	0.0005	0	0	0.0000	0	0
20	0.0818	−6	0	0.0761	−7	1	0.0408	7	23	0.0093	20	31	0.0004	2	3	0.0000	0	0
40	0.3530	20	0	0.3165	22	2	0.1196	64	63	0.0095	35	41	0.0000	0	4	0.0000	0	0
10	0.0898	−5	0	0.0871	−5	1	0.0669	−4	5	0.0355	6	18	0.0091	10	19	0.0002	0	2
20	0.5110	9	0	0.4914	10	0	0.3463	18	5	0.1360	60	59	0.0137	35	41	0.0000	0	0
40	0.9301	−5	0	0.9109	−6	0	0.7033	−42	−39	0.2357	27	19	0.0060	20	23	0.0000	0	0
10	0.3824	1	0	0.3755	1	1	0.3207	5	2	0.2152	17	15	0.0838	33	36	0.0035	5	9
20	0.9339	−2	0	0.9266	−2	0	0.8531	−10	−6	0.6275	−32	−33	0.2067	31	27	0.0010	3	4
40	0.9998	0	0	0.9996	0	1	0.9941	4	4	0.8989	−9	−6	0.2949	−8	−12	0.0000	0	0
10	0.8710	3	0	0.8676	3	0	0.8375	−1	0	0.7579	−11	−2	0.5658	−13	−4	0.1270	21	20
20	0.9999	0	0	0.9999	0	1	0.9998	0	1	0.9957	2	2	0.9452	0	1	0.2491	5	3
40	1.0000	0	0	1.0000	0	1	1.0000	0	1	1.0000	0	1	0.9988	1	2	0.3271	−6	−8

Table 2. Exact values of $F_R(x)$ and errors $(\times 10^4)$ of approximate values for $p = 10$ and $N = 14, 20, 40$ (Error = exact − approximate)

	$\bar{R}$ 0.0			0.1			0.3			0.5			0.7			0.9		
		Error using	Error using		Error using	Error using		Error using	Error using		Error using	Error using		Error using	Error using		Error using	Error using
N	Prob.	(12)	(10)	Prob.	(12)	(10)	Prob.	(12)	(10)	Prob.	(12)	(10)	Prob.	(12)	(10)	Prob.	(12)	(10)
14	0.0000	0	0	0.0000	0	0	0.0000	0	0	0.0000	0	0	0.0000	0	0	0.0000	0	0
20	0.0000	0	0	0.0000	0	0	0.0000	0	0	0.0000	0	0	0.0000	0	0	0.0000	0	0
40	0.0000	0	0	0.0000	0	0	0.0000	0	0	0.0000	0	0	0.0000	0	0	0.0000	0	0
14	0.0001	0	0	0.0001	0	1	0.0001	0	1	0.0000	0	0	0.0000	0	0	0.0000	0	0
20	0.0014	−2	0	0.0013	−2	1	0.0006	−1	1	0.0001	0	1	0.0000	0	0	0.0000	0	0
40	0.0419	−5	0	0.0365	−5	1	0.0113	3	10	0.0006	2	4	0.0000	0	0	0.0000	0	0
14	0.0085	−3	0	0.0082	−3	1	0.0055	−2	1	0.0022	−1	2	0.0003	0	1	0.0000	0	0
20	0.0753	−1	0	0.0712	−1	1	0.0440	0	4	0.0134	5	11	0.0010	2	4	0.0000	0	0
40	0.6149	0	0	0.5841	1	0	0.3542	4	−3	0.0761	40	41	0.0011	4	6	0.0000	0	0
14	0.1330	3	0	0.1292	3	1	0.1010	3	1	0.0547	4	5	0.0131	4	7	0.0001	0	0
20	0.5436	0	0	0.5310	0	0	0.4297	1	0	0.2383	11	7	0.0489	22	25	0.0001	0	1
40	0.9922	0	0	0.9901	0	1	0.9560	0	2	0.7345	−27	−24	0.1469	23	22	0.0000	0	1
14	0.7187	−16	0	0.7136	−16	0	0.6701	−17	0	0.5636	−15	0	0.3475	0	2	0.0345	7	8
20	0.9884	4	0	0.9875	4	1	0.9780	4	1	0.9386	1	0	0.7557	−12	−8	0.0907	13	13
40	1.0000	0	0	1.0000	0	1	1.0000	0	1	1.0000	0	1	0.9934	3	3	0.1990	4	3

References

Elfving, G. (1947). A simple method of deducing certain distributions connected with multivariate sampling. *Skand. Aktuarietidskrift* **30**, 56–74.

Erdélyi, A. (1954). *Tables of Integral Transforms.* McGraw-Hill, New York.

Fisher, R. A. (1928). The general sampling distribution of the multiple correlation coefficient. *Proc. Roy. Soc. London Ser.A* **121**, 654–673.

Gurland, J. (1968). A relatively simple form of the distribution of the multiple correlation coefficient. *J. Roy. Statist. Soc. Ser.B* **30**, 276–283.

Gurland, J. and Milton, R. (1970). Further consideration of the distribution of the multiple correlation coefficient. *J. Roy. Statist. Soc. Ser.B* **32**, 381–394.

Hodgson, V. (1968). On the sampling distribution of the multiple correlation coefficient (abstract). *Ann. Math. Statist.* **39**, 307.

Kendall, M. G. and Stuart, A. (1969). *Distribution Theory.* Griffin, London.

Lee, Y.-S. (1971). Some results on the sampling distribution of the multiple correlation coefficient. *J. Roy. Statist. Soc. Ser.B* **33**, 117–129.

Ruben, H. (1966). Some new results on the distribution of the sample correlation coefficient. *J. Roy. Statist. Soc. Ser.B* **28**, 513–525.

Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706, U.S.A.

(Received November 1989; accepted January 1991)