NOTES

AN APPROXIMATION TO THE DISTRIBUTION OF SAMPLE CORRELATION COFFICIENT, WHEN THE POPULATION IS NON-NORMAL

By PIJUSH DASGUPTA
Indian Statistical Institute

SUMMARY. This paper presents an approximation to the distribution function of $X = \frac{1}{2}(r+1)$, where r is the sample correlation coefficient, in terms of an incomplete beta integral and Jacobi polynomials. The constants involved in the expression are obtained in terms of the moments of r, which are functions of the bivariate cumulants of the parent population. This approximation coincides with the exact expression if the population is bivariate normal with zero correlation coefficient.

1. An approximation to the sampling distribution of the correlation coefficient (r) in samples from any non-normal population when the population correlation coefficient (ρ) is zero was given by Quensel. Gayen (1951) obtained a more general result when population correlation coefficient is not necessarily zero. He obtained this approximation by starting with a bivariate Gram-Charlier expansion of the joint probability density function of the population and ignoring all joint cumulants of the population above the fourth. An alternative approach is presented in this paper. The probability density function of $X = (r+1)/2$ is expanded in terms of a beta density function and Jacobi polynomials. Similar methods were used by Durbin and Watson (1951) in deriving an approximation for the distribution of a statistic used for testing serial correlation in least square regression and by Roy (1965) in approximating the power of Wilks' likelihood ratio test.

2. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be n pairs of observations drawn at random from a continuous bivariate population with cumulants k_{ij}, $i, j = 1, 2, \ldots$, Let us denote by $f(x)$ the probability density function of $X = \frac{1+r}{2}$ where $r = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2} \times \sqrt{\Sigma(y_i-\bar{y})^2}}$. The quotient $f(x)/\beta(x, a, b)$ where $\beta(x, a, b) = x^{a-1}(1-x)^{b-1}/B(a, b)$, can formally be expanded in an infinite series as

$$f(x) = \beta(x, a, b) \sum_{r=0}^{\infty} a_r J_r(x, a, b) \qquad \ldots (1)$$

where $J_r(x, a, b)$ is the Jacobi polynomial of degree r. These polynomials are orthogonal with respect to the beta density function $\beta(x, a, b)$, so that

$$\int_0^1 J_r(x, a, b) J_s(x, a, b)\, \beta(x, a, b) dx = \delta rs\, k_r(a, b) \qquad \ldots (2)$$

where $k_0(a, b) = 1$

$$k_r(a, b) = \frac{a(a+1)\ldots(a+r-1)\, b(b+1)\ldots(b+r-1)}{r!(2r+a+b-1)(a+b)(a+b+1)\ldots(a+b+r-2)} \quad \text{for } r = 1, 2, \ldots .$$

The polynomials are defined by (for $a, b > 0$)

$$J_0(x, a, b) = 1$$

$$J_r(x, a, b) = \sum_{s=0}^{r} (-1)^s\, C_s(r, a, b) x^s \qquad \text{for } r = 1, 2, \ldots$$

where

$$C_0(r, a, b) = a(a+1) \ldots (a+r-1)/r!$$

$$C_\nu(r, a, b) = (r+a+b-1)(r+a+b) \ldots (r+a+b+\nu-2)/\nu!$$
$$\times(a+\nu)(a+\nu+1) \ldots (a+r-1)/(r-\nu)! \quad \nu = 1, 2, \ldots r-1$$

$$C_r(r, a, b) = (r+a+b-1)(r+a+b) \ldots (2r+a+b-2)/r!$$

The parameters a_r in (1) can be computed by multiplying both sides of (1) by $J_s(x, a, b)$ and integrating over x from 0 to 1, when we get

$$a_r = \int_0^1 J_r f(x)dx/k_r$$

formally, by virtue of the orthogonality property (2).

Retaining only first five terms of the expansion we get

$$f(x) = \beta(x, a, b) \sum_{r=0}^{4} a_r J_r(x; a, b) \qquad \ldots (3)$$

and integrating (3) the cumulative destribution function of x is obtained as

$$F(x) = B(x, a, b) - \beta(x, a+1, b+1) \sum_{r=1}^{4} a_r^* J_{r-1}(x, a+1, b+1) \qquad \ldots (4)$$

where $B(x, a, b) = \int_0^x \beta(t, a, b)dt$ and $a_r^* = \dfrac{a_r \cdot ab}{r(a+b)(a+b+1)}$.

Let us write μ_r for the r-th moment of r about origin. We now choose a and b to make $a_1 = a_2 = 0$ This gives

$$a = \frac{2(1+\mu_1)^2-(1+\mu_2)(1+2\mu_1++\mu_2)}{2(1+2\mu_1+\mu_2)-2(1+\mu_1)^2}, \quad b = -\frac{(1-\mu_1)(1-\mu_2)}{2(1+2\mu_1+\mu_2)-2(1+\mu_1)^2}.$$

Writing $C = a+b$ we get

$$a_1^* = 0$$

$$a_2^* = 0$$

$$a_3^* = \frac{C+5}{(b+1)(b+2)}\left[\frac{a}{3} - (C+2)m_1 + \frac{(C+2)(C+3)}{a+1}m_2 - \frac{(C+2)(C+3)(C+4)}{3(a+1)(a+2)}m_3\right]$$

and

$$a_4^* = \frac{(C+7)(C+2)}{(b+1)(b+2)(b+3)}\left[\frac{a}{4} - (C+3)m_1 + \frac{3(C+3)(C+4)}{2(a+1)}m_2 - \frac{(C+3)(C+4)(C+5)}{(a+1)(a+2)}m_3\right.$$
$$\left. + \frac{(C+3)(C+4)(C+5)(C+6)}{4(a+1)(a+2)(a+3)}m_4\right] \qquad \ldots (5)$$

where $m_g = \dfrac{1}{2^g}\sum_{i=0}^{4}\binom{\gamma}{i}^{\mu}_{i}, \qquad g = 1, 2, 3, 4$

The expressions for $\mu_k (k = 1, 2, 3, 4)$ are given by Cook (1951) in terms of the population cumulants to order n^{-3} and are not reproduced here. Thus knowing the first four moments of r expressions for the density function and cumulative distribution function can be obtained using (3) and (4) respectively.

3. When parent population is bivariate normal $N(0, 0, 1, 1, \rho)$, $k_{10} = k_{01} = 0$, $k_{20} = k_{02} = 1$ and $k_{11} = \rho$, all the other cumulants are zero. Here the first four moments of r reduce to (to order n^{-2})

$$\mu_1 = \rho\left\{1 - \frac{1}{2n} - \frac{3}{8n^2} + \rho^2\left(\frac{1}{2n} - \frac{3}{4n^2}\right) + \frac{\rho^4}{8n^2}\right\}$$

$$\mu_2 = \rho^2\left\{1 - \frac{3}{n} + \frac{3}{n^2} + \rho^2\left(\frac{2}{n} - \frac{12}{n^2}\right) + \frac{1}{\rho^2}\left(\frac{1}{n} + \frac{1}{n^2}\right) + \rho^4\,\frac{8}{n^2}\right\}$$

$$\mu_3 = \rho^3\left\{\frac{1}{\rho^2}\left(\frac{3}{n} - \frac{9}{2n^2}\right) + 1 - \frac{15}{2n} + \frac{261}{8n^2} + \rho^2\left(\frac{9}{2n} - \frac{225}{4n^2}\right) + \frac{225}{8n^2}\,\rho^4\right\}$$

$$\mu_4 = \rho^4\left\{\frac{1}{\rho^4}\,\frac{3}{n^2} + \frac{1}{\rho^2}\left(\frac{6}{n} - \frac{36}{n^2}\right) + 1 - \frac{14}{n} + \frac{129}{n^2} + \rho^4\left(\frac{8}{n} - \frac{168}{n^2}\right) + \frac{72}{n^2}\,\rho^4\right\}.$$

We present below a few values of the distribution function of r as given in David's (1938) table and the corresponding approximate values using (4), when the parent population is bivariate normal.

TABLE 1

			prob $(r \leq y \mid \rho)$		
n	ρ	y	exact	Jacobi approximation	% error
10	0.2	0.2	.4850	.5001	4.8
10	0.2	0.6	.9011	.8905	1.2
10	0.4	0.2	.2494	.2550	2.2
10	0.4	0.6	.7480	.7607	1.7
25	0.2	0.2	.4917	.5040	2.5
25	0.2	0.6	.9882	.9826	.6
25	0.4	0.2	.1386	.1205	13.1
25	0.4	0.6	.8910	.8823	1.0
50	0.2	0.2	.4912	.5025	1.7
50	0.2	0.6	.9995	.9989	0.1
100	0.2	0.2	.4959	.5017	1.2

4. As the explicit expression of the probability integral of Gayen's approximation is not available, the approximation developed here, is not compared numerically with Gayen's result. However, Table 1 shows that the agreement between the present approximation and the exact values, when the population is bivariate normal—the only case where the distribution is known exactly, is quite good. The approximation coincides with the exact distribution if the population is normal with zero correlation coefficient.

ACKNOWLEDGEMENT

I am grateful to Dr. J. Roy for his suggestions.

REFERENCES

COOK, M. B. (1951) : Two applications of bivariate k statistics. *Biometrika*, 38, 368-376.

DAVID, F. N. (1938) : *Tables of the Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples*, Cambridge University Press.

DURBIN, J. and WATSON, G. S. (1951) : Testing for serial correlation in least squares regression II. *Biometrika*, 38, 159-178.

GAYEN, A. K. (1951) : The frequency distribution of the product moment correlation coefficient in random samples of any size drawn from non-normal universes. *Biometrika*, 38, 219-247.

QUENSEL, C. E. (1938) : The distribution of the second moment and of the correlation coefficient in samples from populations of type—Au Lunds University Aress. *N. F. Adv.*, 2, Bd. 34, Nr. 4.

ROY, J. (1965) : Power of the likelihood ratio test used in analysis of dispersion. Paper presented at the International Symposium of Multivariate Analysis at Dayton.

SZEGO, G. (1939) : Orthogonal polynomials. *Amer. Math. Soc.*, Colloquium Publication, 13.

Paper received : November, 1966.

Revised : August, 1967.