

Distribution of the correlation coefficient for the class of bivariate elliptical models*

Mir M. ALI and Anwarul H. JOARDER

The University of Western Ontario

Key words and phrases: Pearson product-moment correlation, correlation coefficient, product-moment correlation, distribution of correlation coefficient, elliptically contoured model, bivariate elliptical model.

AMS 1985 subject classifications: Primary 62H20, 62E15, 62H10.

ABSTRACT

We consider n pairs of random variables $(X_{11}, X_{21}), (X_{12}, X_{22}), \dots, (X_{1n}, X_{2n})$ having a bivariate elliptically contoured density of the form

$$K(n)|\Lambda|^{-n/2}g\left\{\sum_1^n(\mathbf{x}_{1j}-\boldsymbol{\theta}_1, \mathbf{x}_{2j}-\boldsymbol{\theta}_2)\Lambda^{-1}(\mathbf{x}_{1j}-\boldsymbol{\theta}_1, \mathbf{x}_{2j}-\boldsymbol{\theta}_2)^T\right\},$$

where $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ are location parameters and $\Lambda = ((\lambda_{ik}))$ is a 2×2 symmetric positive definite matrix of scale parameters. The exact distribution of the Pearson product-moment correlation coefficient between X_1 and X_2 is obtained. The usual case when a sample of size n is drawn from a bivariate normal population is a special case of the abovementioned model.

RÉSUMÉ

Nous considérons n paires de variables aléatoires $(X_{11}, X_{21}), (X_{12}, X_{22}), \dots, (X_{1n}, X_{2n})$ ayant une densité conjointe de la forme

$$K(n)|\Lambda|^{-n/2}g\left\{\sum_1^n(\mathbf{x}_{1j}-\boldsymbol{\theta}_1, \mathbf{x}_{2j}-\boldsymbol{\theta}_2)\Lambda^{-1}(\mathbf{x}_{1j}-\boldsymbol{\theta}_1, \mathbf{x}_{2j}-\boldsymbol{\theta}_2)^T\right\},$$

où $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ sont des paramètres de position et où $\Lambda = ((\lambda_{ik}))$ est une matrice 2×2 de paramètres de dispersion. On obtient la distribution exacte du coefficient de corrélation de Pearson entre X_1 et X_2 . La situation où un échantillon de taille n provient d'une loi binormale représente un cas particulier de celle considérée ici.

1. INTRODUCTION

Consider n pairs of identically distributed (not necessarily independent) random variables (X_{1j}, X_{2j}) for $j = 1, 2, \dots, n$ as realizations of the pair (X_1, X_2) . The statistic commonly used in the study of the degree or strength of the relationship between X_1

*This research was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada, grant no. OGP0002974, to the first author.

and X_2 is known as the Pearson product-moment correlation, usually designated by r and given by

$$r = \frac{\sum_{j=1}^n (X_{1j} - \bar{X}_1)(X_{2j} - \bar{X}_2)}{\left(\sum_{j=1}^n (X_{1j} - \bar{X}_1)^2 \sum_{j=1}^n (X_{2j} - \bar{X}_2)^2 \right)^{\frac{1}{2}}}.$$

The statistic r has a long history and we simply refer to Pearson (1920) and Kruskal (1958), among others, for a discussion on its appropriateness as a suitable measure of association even in nonnormal cases. The quantity

$$\rho = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1)\text{Var}(X_2)}}$$

is called the population correlation coefficient. The properties of ρ for nonnormal models are discussed in most standard texts, e.g., Cramer (1946, p. 275).

Fisher (1915) obtained the exact sampling distribution of r for a random sample drawn from a bivariate normal population. The present work was primarily motivated by the work of Muirhead (1982), who studied the asymptotic distribution of r based on a random sample from a bivariate elliptical distribution. The elliptical distributions have been studied by several authors, among whom we mention Dawid (1977), Chmielewski (1981), Fraser and Ng (1980), Kariya (1981), Eaton (1983), and Anderson, Fang, and Hsu (1986). Gayen (1951) and Devlin, Gnanadesikan, and Kettenring (1976) studied the robustness of the distribution of r . The reader is referred to Kowalski (1972) for a historical survey of the literature on robustness studies.

In this paper we consider the elliptical distribution (Sutradhar and Ali 1989) of n two-dimensional random vectors $X_1, X_2, \dots, X_n$ having a joint p.d.f. of the form

$$f(\mathbf{x}_1^T, \mathbf{x}_2^T, \dots, \mathbf{x}_n^T; \boldsymbol{\theta}, \boldsymbol{\Lambda}) = K(n) |\boldsymbol{\Lambda}|^{-n/2} g \left(\sum_{j=1}^n (\mathbf{x}_j - \boldsymbol{\theta})^T \boldsymbol{\Lambda}^{-1} (\mathbf{x}_j - \boldsymbol{\theta}) \right), \quad (1.1)$$

where $K(n)$ is a suitable normalizing constant, $\mathbf{x}_j^T = (x_{1j}, x_{2j})$ is a two-dimensional row vector, $\boldsymbol{\theta}' = (\theta_1, \theta_2)$ is a location parameter, and

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix}$$

is a 2×2 positive definite matrix of scale parameters.

We derive the exact distribution of r based on the above model and show that the distribution of r depends only on n and the parameter $\rho = \lambda_{12}/\sqrt{\lambda_{11}\lambda_{22}}$ and does not depend on the form of $g(\cdot)$. In particular, the case where $X_1, X_2, \dots, X_n$ is a sample from the bivariate normal population with parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Lambda}$ belongs to the class given by (1.1). We remark that if the second moments of X_j for $j = 1, 2, \dots, n$ in the model (1.1) exist, say

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix},$$

then it follows from Ali and Ponnappalli (1990, p. 174) that the matrix $\boldsymbol{\Lambda}$ determines the matrix $\boldsymbol{\Sigma}$ up to a scale factor c^2 . More precisely, $\boldsymbol{\Sigma} = c^2 \boldsymbol{\Lambda}$, where c^2 is an arbitrary positive constant. It then follows that in our case, $\lambda_{12}/\sqrt{\lambda_{11}\lambda_{22}} = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}} = \rho$ is also the population correlation coefficient for this bivariate elliptical model.

For the model (1.1), it is shown by Anderson, Fang, and Hsu (1986, p. 58, Example 1) that the maximum-likelihood estimators of σ_{11} , σ_{22} , and σ_{12} are given by

$$\hat{\sigma}_{11} = \frac{2}{y_g} \sum_1^n (X_{1j} - \bar{X}_1)^2, \quad \hat{\sigma}_{22} = \frac{2}{y_g} \sum_1^n (X_{2j} - \bar{X}_2)^2,$$

and

$$\hat{\sigma}_{12} = \frac{2}{y_g} \sum_1^n (X_{1j} - \bar{X}_1)(X_{2j} - \bar{X}_2),$$

where y_g is the maximum of $y^{n/2}g(y)$. It follows that the maximum-likelihood estimator of ρ is then given by

$$\hat{\rho} = \frac{\hat{\sigma}_{12}}{(\hat{\sigma}_{11}\hat{\sigma}_{22})^{1/2}} = r.$$

An important problem in the area of stock-market analysis is the study of the performance of stocks of some selected firms, say, (1) Bell Canada, (2) Bank of Montreal, relative to the performance of all stocks trading in the category of blue-chip stocks (hereinafter called the "market" for short). Let y_{ij} be the monthly return on the i th stock in the j th month, and m_j denote the weighted average of these returns during the j th month for the entire market. A regression model would then be

$$y_{ij} = \alpha_i + \beta_i m_j + \epsilon_{ij}$$

for $i = 1, 2$ and $j = 1, 2, \dots, n$. We now assume that $\mathcal{E}(\epsilon_{ij}) = 0$, $\mathcal{E}(\epsilon_{ij}^2) = \sigma^2 \lambda_{11}$, $\mathcal{E}(\epsilon_{2j}^2) = \sigma^2 \lambda_{22}$, $\mathcal{E}(\epsilon_{1j}, \epsilon_{2j}) = \sigma^2 \lambda_{12}$, and $\mathcal{E}(\epsilon_{ij}\epsilon_{lk}) = 0$ with $j \neq k$, for all $i = 1, 2$, $l = 1, 2$, and $j, k = 1, 2, \dots, n$. Furthermore, we assume that for a given σ , the error vectors $\epsilon_j = (\epsilon_{1j}, \epsilon_{2j})^T$, $j = 1, 2, \dots, n$, are identically and independently distributed according to $N_2(0, \sigma^2 \Lambda)$. We also assume that σ , a parameter representing monthly fluctuations, is a random variable having an inverted gamma distribution (cf. Zellner 1976). Then it can be shown that the density of $(\epsilon_{ij} : i = 1, 2, j = 1, 2, \dots, n)$ is given by a multivariate t -distribution which belongs to the model (1.1). For further details the reader is referred to Sutradhar and Ali (1986).

Mandelbrot (1963) and Fama (1965) suggested that observed "fat tails" of the rates of return of common stock could be studied by using the symmetric stable distributions. Once again these distributions form a subclass of the model (1.1).

2. THE RESULT

THEOREM. Consider the vector random variable $\mathbf{X} = (\mathbf{X}_1^T, \mathbf{X}_2^T, \dots, \mathbf{X}_n^T)$, where $\mathbf{X}_j^T = (X_{1j}, X_{2j})$ is a two-dimensional random variable for $j = 1, 2, \dots, n$, having a p.d.f. given by (1.1). Then the p.d.f. of r is given by

$$\begin{aligned} f(r) &= \frac{n-2}{\pi} (1-\rho^2)^{(n-1)/2} (1-r^2)^{(n-4)/2} \times \int_0^1 \frac{x^{n-2} dx}{(1-\rho r x)^{n-1} \sqrt{1-x^2}} \\ &= \frac{2^{n-3}}{(n-3)! \pi} (1-\rho^2)^{(n-1)/2} (1-r^2)^{(n-4)/2} \times \sum_{\alpha=0}^{\infty} \frac{(2\rho r)^\alpha}{\alpha!} \Gamma^2\left(\frac{n+\alpha-1}{2}\right) \end{aligned} \quad (2.1)$$

for $-1 < r < 1$, where $\rho + \lambda_{12}/\sqrt{\lambda_{11}\lambda_{22}}$.

Proof. The transformation $Z_{ij} = (X_{ij} - \theta_i)/\sqrt{\lambda_{ii}}$ for $i = 1, 2$ and $j = 1, 2, \dots, n$, applied to (1.1) with $\mathbf{Z}_j^T = (Z_{1j}, Z_{2j})$, gives

$$f(z'_1, z'_2, \dots, z'_n) = K(n) \left| \begin{array}{cc} 1 & \rho \\ \rho & 1 \end{array} \right|^{-n/2} g \left(\sum_1^n z'_j \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} z_j \right). \quad (2.2)$$

It is readily verified that r is also given by $r = a_{12}/\sqrt{a_{11}a_{22}}$, where $a_{12} = \sum_1^n (z_{1j} - \bar{z}_1)(z_{2j} - \bar{z}_2)$, $a_{11} = \sum_1^n (z_{1j} - \bar{z}_1)^2$, and $a_{22} = \sum_1^n (z_{2j} - \bar{z}_2)^2$.

For the special model (2.2) it follows from Theorem 2.1 Equation (2.2) of Sutradhar and Ali (1989) that the joint p.d.f. of a_{11} , a_{12} , and a_{22} is given by

$$f(a_{11}, a_{12}, a_{22}) = K(n) g_{(n-1,2)} \left(\text{tr} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \right) \\ \times \left| \begin{array}{cc} 1 & \rho \\ \rho & 1 \end{array} \right|^{-(n-1)/2} \left| \begin{array}{cc} a_{11} & a_{12} \\ a_{21} & a_{22} \end{array} \right|^{(n-4)/2},$$

where $g_{(n-1,2)}$ is determined by n and the form of $g(\cdot)$, and $K(n)$ denotes a generic normalizing constant. The transformation $a_{11} = s_1^2$, $a_{12} = rs_1s_2$, and $a_{22} = s_2^2$, having the Jacobian $(2s_1s_2)^2$, yields the p.d.f. of s_1 , s_2 , and r as

$$f(s_1, s_2, r) = K(n) (1 - \rho^2)^{-(n-1)/2} (1 - r^2)^{(n-4)/2} (s_1s_2)^{n-2} \\ \times g_{(n-1,2)} \left(\frac{1}{1 - \rho^2} (s_1^2 - 2\rho rs_1s_2 + s_2^2) \right),$$

where $0 < s_1 < \infty$, $0 < s_2 < \infty$, $-1 < r < 1$.

Next, the polar transformation $s_1 = s \cos \theta$, $s_2 = s \sin \theta$, and $r = r$ followed by integration over s gives the density of r and θ as

$$f(r, \theta) = K(n) (1 - \rho^2)^{-(n-1)/2} (1 - r^2)^{(n-4)/2} (\sin 2\theta)^{n-2} \\ \times \int_0^\infty s^{2n-3} g_{(n-1,2)} \left(\frac{1 - \rho r \sin 2\theta}{1 - \rho^2} s^2 \right) ds.$$

To evaluate the above integral, let

$$\frac{1 - \rho r \sin 2\theta}{1 - \rho^2} s^2 = u,$$

so that

$$\frac{1 - \rho r \sin 2\theta}{1 - \rho^2} 2s ds = du, \quad \text{for } 0 < \theta < \frac{\pi}{2}.$$

A simple calculation shows that

$$f(r, \theta) = K(n) (1 - \rho^2)^{(n-1)/2} (1 - r^2)^{(n-4)/2} \frac{(\sin 2\theta)^{n-2}}{(1 - \rho r \sin 2\theta)^{n-1}} \\ \times \int_0^\infty u^{n-2} g_{(n-1,2)}(u) du, \quad 0 < \theta < \frac{\pi}{2}.$$

Now $f(r, \theta)$ is an even function of θ centered at $\theta = \pi/4$, so that

$$f(r) = K(n) (1 - \rho^2)^{(n-1)/2} (1 - r^2)^{(n-4)/2} \int_0^{\pi/4} \frac{(\sin 2\theta)^{n-2}}{(1 - \rho r \sin 2\theta)^{n-1}} d\theta.$$

Finally, set $\sin 2\theta = x$, so that

$$f(r) = K(n)(1 - \rho^2)^{(n-1)/2}(1 - r^2)^{(n-4)/2} \int_0^1 \frac{x^{n-2}}{(1 - \rho r x)^{n-1}} \frac{dx}{\sqrt{1 - x^2}}.$$

Comparing this expression for $f(r)$ with that of Cramer [1946, p. 398, Equation (29.7.2)], we immediately obtain $K(n)$, so that $f(r)$ is given by (2.1), and hence the theorem is proved. Q.E.D.

A qualitative proof of the above theorem by the use of invariance principle is as follows: Let $(X_{ij} - \theta_i)/\sqrt{\lambda_{ii}} = \sqrt{1 + \rho}V_{1j} + (-1)^{i+1}\sqrt{1 - \rho}V_{2j}$ for $i = 1, 2, j = 1, 2, \dots, n$. Then from (1.1) it is readily verified that the $2n$ -dimensional random variable $(V_{11}, V_{22}, \dots, V_{1n}, V_{21}, \dots, V_{2n})$ is spherically distributed. It follows that $U_{ij} = V_{ij}/\|V\|$ (where $\|V\|^2 = \sum_{i=1}^2 \sum_{j=1}^n V_{ij}^2$) for $i = 1, 2, j = 1, 2, \dots, n$, is uniformly distributed on the surface of $2n$ -dimensional unit sphere with center at the origin. Thus the joint density of $U_{11}, U_{22}, \dots, U_{1n}, U_{21}, \dots, U_{2n}$ does not depend on $g(\cdot)$. It is easy to check that r is a function of U_{ij} ($i = 1, 2, j = 1, 2, \dots, n$), namely

$$r = \frac{(1 + \rho) \sum_1^n (U_{1j} - \bar{U}_1)^2 - (1 - \rho) \sum_1^n (U_{2j} - \bar{U}_2)^2}{\left[\sum_1^n \{ \sqrt{1 + \rho}(U_{1j} - \bar{U}_1) + \sqrt{1 - \rho}(U_{2j} - \bar{U}_2) \}^2 \sum_1^n \{ \sqrt{1 + \rho}(U_{1j} - \bar{U}_1) - \sqrt{1 - \rho}(U_{2j} - \bar{U}_2) \}^2 \right]^{\frac{1}{2}}},$$

and it follows that the density of r does not depend on $g(\cdot)$, and hence the theorem is proved.

ACKNOWLEDGEMENT

The second author, who is on leave from the Jahangirnagar University, Bangladesh, is grateful to the Canadian International Development Agency for financial support in the form of a commonwealth scholarship. The authors take this opportunity to thank the referees and the editor for their valuable comments and suggestions leading to substantial improvement of the paper.

REFERENCES

- Ali, M.M., and Ponnappalli, R. (1990). An optimal property of the Gauss-Markoff estimator. *J. Multivariate Anal.*, 32, 171-176.
- Anderson, T.W., Fang, K., and Hsu, H. (1986). Maximum likelihood estimates and likelihood-ratio criterion for multivariate elliptically contoured distributions. *Canad. J. Statist.*, 14, 55-59.
- Chmielewski, M.A. (1981). Elliptically symmetric distributions: A review and bibliography. *Internat. Statist. Rev.*, 49, 67-74.
- Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton Univ. Press, Princeton, N.J.
- Dawid, A.P. (1977). Spherical matrix distributions and a multivariate model. *J. Roy. Statist. Soc. Ser. B*, 39, 254-261.
- Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1976). Some multivariate applications of elliptical distributions. *Essays in Probability and Statistics* (S. Iyengar, ed.), Sinko Tsusho, Tokyo, 365-395.
- Eaton, M.L. (1983). *Multivariate Statistics: A Vector Space Approach*. Wiley, New York, Chapter 7.
- Fama, E.F. (1965). The behaviour of stock market prices. *J. Business*, 38, 34-105.
- Fisher, R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10, 507-521.
- Fraser, D.A.S., and Ng, K.W. (1980). Multivariate regression analysis with spherical error. *Multivariate Analysis* 5 (P.R. Krishnaiah, ed.), North Holland, New York, 369-386.
- 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.

- Kariya, T. (1981). Robustness of multivariate tests. *Ann. Statist.*, 9, 1267-1275.
- Kowalski, D.J. (1972). On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. *Appl. Statist.*, 21, 1-12.
- Kruskal, W.H. (1958). Ordinal measures of association. *J. Amer. Statist. Assoc.*, 53, 814-861.
- Mandelbrot, B. (1963). The variation of certain speculative prices. *J. Business*, 36, 394-419.
- Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York, Chapter 5.
- Pearson, K. (1920). Notes on history of correlations. *Biometrika*, 13, 25-45.
- Sutradhar, B.C., and Ali, M.M. (1986). Estimation of the parameters of a regression model with a multivariate t error variable. *Comm. Statist. A Theory Methods*, 15(2), 429-450.
- Sutradhar, B.C., and Ali, M.M. (1989). A generalization of the Wishart distribution for the elliptical model and its moments for the multivariate t model. *J. Multivariate Anal.*, 29, 155-162.
- Zellner, A. (1976). Bayesian and non-Bayesian analysis of the regression model with multivariate Student t -error term. *J. Amer. Statist. Assoc.*, 71, 400-405.

Received 27 August 1990

Revised 26 February 1991

Accepted 8 April 1991

Department of Statistics and Actuarial Sciences

The University of Western Ontario

London, ON N6A 5B9