Biometrika (1983), **70**, 2, *pp.* 327–32
Printed in Great Britain

A correlation coefficient for circular data

By N. I. FISHER

CSIRO Division of Mathematics and Statistics, Lindfield, N.S.W., Australia

AND A. J. LEE

Department of Mathematics, University of Auckland, Auckland, New Zealand

Summary

A coefficient to measure association between angular variables is discussed, its asymptotic distribution found, and its properties developed. Comparisons with other statistics in current use are made, and some examples given.

Some key words: Circular correlation; Directional data; Rank correlation; T-linear association; T-monotone association; U-statistic.

1. Introduction

Let Θ and Φ be two angular random variables with joint distribution on the surface of a torus. A natural way of defining complete dependence of Θ and Φ, corresponding to a linear relationship between two real random variables, is

$$\Theta \equiv \Phi + \alpha_0 \mod(2\pi), \quad \text{positive association}, \tag{1·1}$$

$$\Theta \equiv -\Phi + \alpha_0 \mod(2\pi), \quad \text{negative association}, \tag{1·2}$$

for some arbitrary fixed direction α_0. We shall refer to such dependences as toroidal-linear, or T-linear.

Much of the recent research on T-linear dependence has been concerned with detecting the existence of this sort of association, without discriminating between positive and negative association; thus, the measures proposed to estimate T-linear association have necessarily taken values in the range $[0, 1]$. Jupp & Mardia (1980) give a comprehensive review. However, Rivest (1982) has proposed a signed measure of T-linear association taking values in the interval $[-1, 1]$ which permits discrimination between (1·1) and (1·2). Rivest's measure ρ_c^* is the smaller eigenvalue of the cross-product matrix $E(uv')$, where $u' = (\cos\Theta, \sin\Theta)$ and $v' = (\cos\Phi, \sin\Phi)$.

In the present paper we study a correlation coefficient ρ_T which is also a signed measure of T-linear assocation, and which is a natural analogue of the usual product moment correlation of two linear variates. In §2, the definition and properties of ρ_T are presented, and ρ_T is evaluated for two parametric families. In §3, a suitable sample estimate $\hat{\rho}_T$ is proposed, and appropriate large sample theory developed. Some numerical examples are considered in §4.

Corresponding to monotone association between two real random variables, it is possible to define a notion of toroidal-monotone, or T-monotone, association between two angular random variables. Fisher & Lee (1982) proposed a model of T-monotone association and introduced a U-statistic analogous to Kendall's tau to estimate it. An analogue of Spearman's rho was also proposed; its connexion with $\hat{\rho}_T$ is discussed in §5.

2. Definition of the correlation coefficient

We propose to measure association between two circular variates Θ and Φ by measuring separately the degree to which Θ can be predicted from Φ using one of the relationships (1·1) or (1·2), and constructing a measure by taking the difference of these measures. This seems reasonable in that a relationship well predicted by (1·1) will be poorly predicted by (1·2) and vice versa. The result is a signed correlation coefficient which permits discrimination between relationships of the form (1·1) and (1·2).

Specifically, consider the quantities

$$S\pm(\alpha) = E[\{\cos\Theta-\cos(\alpha\pm\Phi)\}^2]+E[\{\sin\Theta-\sin(\alpha\pm\Phi)\}^2].$$

The minima of these expressions as α varies are given by

$$S_+ = 2\{1-R(\Theta-\Phi)\},\quad S_- = 2\{1-R(\Theta+\Phi)\},$$

where $R^2(\Psi)\equiv\{E(\cos\Psi)\}^2+\{E(\sin\Psi)\}^2$ is the square of the mean resultant length of a circular variate Ψ. Thus S_+ and S_- measure the extent to which Θ may be predicted from Φ using (1·1) and (1·2), and

$$S_- - S_+ = 2\{R(\Theta-\Phi)-R(\Theta+\Phi)\}. \qquad (2·1)$$

When (1·1) is true, we can expect the first term of this to be unity, and the second small, while if (1·2) is true the reverse obtains.

Rivest's measure is based on a normed version of (2·1); for the properties of this measure, see Rivest (1982).

However, considerable statistical and mathematical advantages may be gained by considering instead a measure based on

$$D = 2\{R^2(\Theta-\Phi)-R^2(\Theta+\Phi)\}$$

and we now turn to a discussion of such a measure.

Suppose that (Θ_1,Φ_1) and (Θ_2,Φ_2) are independently distributed as (Θ,Φ). Then we may write

$$D = E\{2\sin(\Theta_1-\Theta_2)\sin(\Phi_1-\Phi_2)\},$$

whence

$$D^2 \leqslant \{1-R^2(2\Theta)\}\{1-R^2(2\Phi)\}.$$

It is thus natural to norm the quantity D and define a signed measure of the predictability of Θ from Φ by

$$\rho_T = D/[\{1-R^2(2\Theta)\}\{1-R^2(2\Phi)\}]^{\frac{1}{2}}$$
$$= \frac{E[\{\sin(\Theta_1-\Theta_2)\sin(\Phi_1-\Phi_2)\}]}{[E\{\sin^2(\Theta_1-\Theta_2)\}E\{\sin^2(\Phi_1-\Phi_2)\}]^{\frac{1}{2}}}. \qquad (2·2)$$

Note that the ordinary correlation coefficient $\rho(X, Y)$ for two linear variables X and Y has the alternative form

$$E\{(X_1-X_2)(Y_1-Y_2)\}/[E\{(X_1-X_2)^2\}E\{(Y_1-Y_2)^2\}]^{\frac{1}{2}}.$$

The measure ρ_T has the following properties:

(i) $-1\leqslant\rho_T\leqslant 1$;

(ii) $\rho_T = 1$ if and only if Θ and Φ are related by (1·1), and $\rho_T = -1$ if and only if (1·2) holds;

(iii) ρ_T is invariant under choice of origin for Θ and Φ, and reflection of one of Θ and Φ changes the sign of ρ_T but not its magnitude;

(iv) if Θ and Φ are independent then $\rho_T = 0$;

(v) if the distributions of Θ and Φ are each unimodal and highly concentrated, $\rho_T(\Theta, \Phi) \doteqdot \rho(\Theta, \Phi)$.

We note that ρ_T has properties (ii) and (v), which are not completely satisfied by Rivest's coefficient ρ_c^*.

Another signed measure of circular correlation was proposed by Thompson (1975), based on half-angles $\frac{1}{2}\Theta$ and $\frac{1}{2}\Phi$ instead of Θ and Φ. However, Thompson's measure has the drawback that it is not necessarily 1 under the model (1·1), or -1 under the model (1·2).

We now present examples of ρ_T computed for specific parametric families of densities.

Example 1. Let X, Y have a bivariate normal distribution with zero mean vector, variances σ_1^2, σ_2^2 and correlation ρ. Define

$$\Theta \equiv X \bmod (2\pi), \quad \Phi \equiv Y \bmod (2\pi),$$

so that Θ and Φ have a wrapped bivariate normal distribution. Then, since $\cos\theta$ and $\sin\phi$ are periodic with period 2π, we have

$$E(\cos\Theta \sin\Phi) = E(\cos X \sin Y),$$

and similarly for the other expressions involved in the numerator of ρ_T. We obtain

$$E(\sin\Theta \sin\Phi) = \exp\{-\tfrac{1}{2}(\sigma_1^2+\sigma_2^2)\} \sinh(\rho\sigma_1\sigma_2),$$

$$E(\cos\Theta \cos\Phi) = \exp\{-\tfrac{1}{2}(\sigma_1^2+\sigma_2^2)\} \cosh(\rho\sigma_1\sigma_2),$$

$$E(\sin\Theta \cos\Phi) = E(\cos\Theta \sin\Phi) = 0,$$

and so

$$D = 2\exp\{-(\sigma_1^2+\sigma_2^2)\} \sinh(2\rho\sigma_1\sigma_2).$$

Since for the wrapped normal $R(2\Theta) = \exp(-2\sigma_1^2)$, $R(2\Phi) = \exp(-2\sigma_2^2)$, we obtain finally

$$\rho_T = \sinh(2\rho\sigma_1\sigma_2)/\{\sinh(2\sigma_1^2)\sinh(2\sigma_2^2)\}^{\frac{1}{2}},$$

which is well approximated by the linear correlation ρ for σ_1, σ_2 small.

Example 2. Consider the family of bivariate densities (Wehrly & Johnson, 1980)

$$f_-(\theta, \phi) = g(\theta-\phi)/(2\pi), \quad f_+(\theta, \phi) = g(\theta+\phi)/(2\pi), \tag{2·3}$$

where g is a density on the circle with mean resultant length R. Then $\rho_T = R^2$ for the first family (2·3) and $-R^2$ for the second. The case of uniform g corresponds to $\rho_T = 0$ while as the resultant length of the mean direction of g increases to 1 the correlation ρ_T increases to 1 in the first case and decreases to -1 in the second.

3. Definition and properties of the estimator

Given a random sample $p_i = (\theta_i, \phi_i)$ $(i = 1, \ldots, n)$ from a distribution on the torus, a natural estimator of ρ_T suggested by (2·2) is

$$\hat{\rho}_T = \frac{\Sigma \sin(\theta_i-\theta_j)\sin(\phi_i-\phi_j)}{\{\Sigma \sin^2(\theta_i-\theta_j)\}^{\frac{1}{2}}\{\Sigma \sin^2(\phi_i-\phi_j)\}^{\frac{1}{2}}},$$

where each summation is over the range $1 \leqslant i < j \leqslant n$. By defining

$$K^{(1)}(P_1, P_2) = \sin(\theta_1 - \theta_2)\sin(\phi_1 - \phi_2),$$

$$K^{(2)}(P_1, P_2) = \sin^2(\theta_1 - \theta_2), \quad K^{(3)}(P_1, P_2) = \sin^2(\phi_1 - \phi_2),$$

and letting $U_n^{(1)}$, $U_n^{(2)}$ and $U_n^{(3)}$ be the U-statistics based on the kernels $K^{(1)}, K^{(2)}, K^{(3)}$ and the sample $p_1, \ldots, p_n$, we can write

$$\hat{\rho}_T = U_n^{(1)}/(U_n^{(2)}\, U_n^{(3)})^{\frac{1}{2}}.$$

We note that $\hat{\rho}_T$ shares properties (i)–(iii) of ρ_T.

4. Asymptotic properties

From the theory of U-statistics (Hoeffding, 1948), we have that

$$\operatorname{cov}(U_n^{(i)}, U_n^{(j)}) = 4\xi_1(i, j)/n + o(n^{-1}),$$

where $\xi_1(i, j) = \operatorname{cov}\{K^{(i)}(P_1, P_2),\ K^{(j)}(P_2, P_3)\}$.

Let $\mu_i = E\{K^{(i)}(P_1, P_2)\}$; then a Taylor series argument yields

$$E(\hat{\rho}_T) = \rho_T\{1 + (3\eta_{22}/2 + 3\eta_{33}/2 - 2\eta_{12} - 2\eta_{13} + \eta_{23})\} + o(n^{-1}),$$

$$\operatorname{var}(\hat{\rho}_T) = n^{-1}\, \rho_T^2(4\eta_{11} + \eta_{22} + \eta_{33} - 4\eta_{12} - 4\eta_{13} + 2\eta_{23}) + o(n^{-1}),$$

where $\eta_{ij} = \xi_1(i, j)/(\mu_i \mu_j)$ $(i, j = 1, 2, 3)$.

In the special case of independence, these reduce to

$$E(\hat{\rho}_T) = o(n^{-1}), \quad \operatorname{var}(\hat{\rho}_T) = 4\xi_1(1, 1)/(n\mu_2\, \mu_3) + o(n^{-1}),$$

since $\xi_1(1, 2) = \xi_1(1, 3) = \xi_1(2, 3) = 0$, and $\mu_1 = 0$.

For an arbitrary circular variate Ψ, define the trigonometric moments

$$\alpha_p = \alpha_p(\Psi) = E\{\cos(p\Psi)\}, \quad \beta_p = \beta_p(\Psi) = E\{\sin(p\Psi)\},$$

and let $A(\Psi) = \alpha_1^2 + \beta_1^2 + \alpha_2\beta_1^2 - \alpha_1^2\alpha_2 - 2\alpha_1\beta_1\beta_2$. Then

$$4\xi_1(1, 1) = A(\Theta)\, A(\Phi), \quad \mu_2 = \tfrac{1}{2}\{1 - \alpha_2^2(\Theta) - \beta_2^2(\Theta)\}$$

and $\mu_3 = \frac{1}{2}\{1 - \alpha_2^2(\Phi) - \beta_2^2(\Phi)\}$, and since sample trigonometric moments consistently estimate the population moments, we may test the independence hypothesis by using the asymptotically $N(0, 1)$ statistic

$$Z = (n^{\frac{1}{2}}\, \hat{\mu}_1\, \hat{\mu}_2\, \hat{\rho}_T)/\{\hat{A}(\Theta)\, \hat{A}(\Phi)\}^{\frac{1}{2}},$$

where the circumflex denotes the replacement of population quantities by sample quantities. For small sample sizes, a randomization test can be used; for moderate sample sizes, it will be necessary to sample the randomization distribution of Z.

If the mean resultant length of Θ or Φ is zero, for example, if either is *a priori* assumed uniform, then $\xi_1(1, 1) = 0$ and the U-statistic $U_n^{(1)}$ is degenerate. It follows from standard U-statistic theory (Gregory, 1977) that in this case the asymptotic distribution of $n\hat{\rho}_T$ is double exponential with density $\frac{1}{2}e^{-|x|}$.

For the point and interval estimation ρ_T in the nonindependent case, we can use the results of Arvesen (1969). Using his notation, define

$$\hat{\rho}_{n-1}^{(i)} = U_{-i}^{(1)}/\{U_{-i}^{(2)}\, U_{-i}^{(3)}\}^{\frac{1}{2}},$$

where, for example, $U^{(1)}_{-i}$ is the U-statistic $U^{(1)}$ calculated on a sample of size $n-1$ with p_i deleted.

Denoting by $\hat{\rho}_i$ the pseudovalues $n\hat{\rho}_T-(n-1)\,\hat{\rho}^{(i)}_{n-1}$, we may use as a point estimate of ρ_T the jackknifed version $\hat{\rho}_J = n^{-1}\,\Sigma_i\,\hat{\rho}_i$ of $\hat{\rho}_T$ and estimate the standard error of $\hat{\rho}_J$ consistently by $n^{-\frac{1}{2}}s$ where

$$s^2 = (n-1)^{-1}\sum_{i=1}^{n}(\hat{\rho}_i-\hat{\rho}_J)^2.$$

A $(1-\alpha)$ confidence interval is accordingly $\hat{\rho}_J \pm n^{-\frac{1}{2}}s\,z(\tfrac{1}{2}\alpha)$ where $z(\tfrac{1}{2}\alpha)$ is the upper $\tfrac{1}{2}\alpha$ point of the standard normal distribution.

The validity of the above depends on the function $\mu_1/(\mu_2\,\mu_3)^{\frac{1}{2}}$ having bounded derivatives in a neighbourhood of (μ_1,μ_2,μ_3), which will be the case if μ_2 and μ_3 are nonzero, i.e. provided that neither 2Θ nor 2Φ is constant.

5. Examples

We reanalyse the sets of data cited by Fisher & Lee (1982) and illustrate them in Figs 1 and 2.

Fig. 1

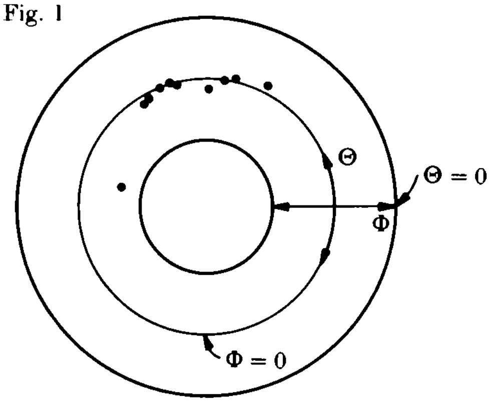

Fig. 2

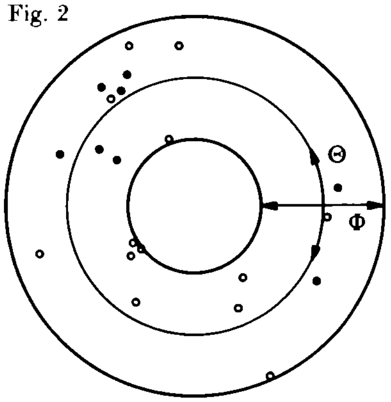

Fig. 1. Peak times of two successive measurements Θ, Φ, of blood pressure (Downs, 1974). Data are projected from the surface of a torus onto a plane normal to the axis of the torus.

Fig. 2. Wind directions at 6.00 am and 12 noon, on 21 successive days (Johnson & Wehrly, 1977; Wehrly & Johnson, 1980). Projection as in Fig. 1, with data on concealed half of torus appearing as open circles.

Example 1. (Downs, 1974.) The peak times for two successive measurements of blood pressure, converted into angles, of 10 medical students were recorded. The estimated value $\hat{\rho}_T$ for these data is 0·965. This very high value is not surprising in view of the plot of the data.

Example 2. (Johnson & Wehrly, 1977; Wehrly & Johnson, 1980.) Analysis of these data by Wehrly & Johnson (1980) suggests that the marginal random variables may reasonably be assumed to have circular uniform distributions. Under this assumption $n\hat{\rho}_T$ is asymptotically unit double exponential under independence, as indicated in the previous section. For these data, $\hat{\rho}_T = 0{\cdot}191$, and $n\hat{\rho}_T = 4{\cdot}011$; the upper 0·025 point of the asymptotic distribution is 2·99, suggesting some degree of T-linearity between the underlying variates.

6. NONPARAMETRIC MEASURES

Fisher & Lee (1982) proposed a general definition of association for two angular variables on the torus, and studied two statistics for assessing this assocation. One of these statistics, $\hat{\Pi}_n$, is an analogue of Spearman's rho and a modification of a statistic proposed by Mardia (1975). Now $\hat{\Pi}_n$ can be written

$$4n^{-2} \sum_{1 \leqslant i < j \leqslant n} \sin\{2\pi(r_i - r_j)/n\} \sin\{2\pi(s_i - s_j)/n\},$$

where $r_1, \ldots, r_n$ and $s_1, \ldots, s_n$ are the ranks referred to an arbitrary origin of $\theta_1, \ldots, \theta_n$ and $\phi_1, \ldots, \phi_n$ respectively. This is identical to $\hat{\rho}_T$ calculated using the circular ranks $2\pi r_i/n$ and $2\pi s_i/n$, for $n > 2$.

Thus $\hat{\rho}_T$ and $\hat{\Pi}_n$ have the same relationship as do Pearson's r and Spearman's rho, in that the latter is obtained from the former by replacing observations by their ranks.

Note that, under independence, the asymptotic distribution of $\hat{\rho}_T$, when the marginals are uniform, coincides with the asymptotic distribution of $\hat{\Pi}_n$.

We are grateful to the referee for a careful reading of the original manuscript.

REFERENCES

ARVESEN, J. A. (1969). Jackknife *U*-statistics. *Ann. Math. Statist.* **40**, 2076–100.

DOWNS, T. D. (1974). Rotational angular correlations. In *Biorhylhms for Human Reproduction*, Eds. M. Ferin, F. Halberg, R. Richart and R. Vande Wiele, Chapter 7, pp. 97–104. New York: Wiley.

FISHER, N. I. & LEE, A. J. (1982). Nonparametric measures of angular-angular association. *Biometrika* **69**, 315–22.

GREGORY, G. G. (1977). Large sample theory for *U*-statistics and tests of fit. *Ann. Statist.* **5**, 110–23.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19**, 293–325.

JOHNSON, R. A. & WEHRLY, T. E. (1977). Measures and models for angular correlation and angular-linear correlation. *J. R. Statist. Soc.* B **39**, 222–9.

JUPP, P. E. & MARDIA, K. V. (1980). A general correlation coefficient for directional data and related regression problems. *Biometrika* **67**, 163–73.

MARDIA, K. V. (1975). Statistics of directional data (with discussion). *J. R. Statist. Soc.* B **37**, 349–93.

RIVEST, L.-P. (1982). Some statistical methods for bivariate circular data. *J. R. Statist. Soc.* B **44**, 81–90.

THOMPSON, J. (1975). Discussion of paper by K. V. Mardia. *J. R. Statist. Soc.* B **37**, 379.

WEHRLY, T. E. & JOHNSON, R. A. (1980). Bivariate models for dependence of angular observations and a related Markov process. *Biometrika* **67**, 255 6.

[*Received September* 1982. *Revised December* 1982]