[4] A. WALD: "The fitting of straight lines if both variables are subject to error," *Annals of Math. Stat.*, Vol. 11 (1940), pp. 284 ff.
[5] T. HAAVELMO: "The probability approach in econometrics," *Econometrica*, Vol. 12 (1944), Supplement.
[6] T. KOOPMANS: *Linear Regression Analysis in Economic Time Series*, Haarlem, 1937.
[7] G. TINTNER: "An application of the variate difference method to multiple regression," *Econometrica*, Vol. 12 (1944), pp. 97 ff.
[8] H. HOTELLING: "Simplified calculation of principal components," *Psychometrica*, Vol. 1 (1936), pp. 27 ff.
[9] T. W. ANDERSON AND M. A. GIRSHICK: "Some extensions of the Wishart distribution," *Annals of Math. Stat.*, Vol. 15 (1944), pp. 354 ff.

NOTE ON THE DISTRIBUTION OF THE SERIAL CORRELATION COEFFICIENT[1]

By WILLIAM G. MADOW

Bureau of the Census

The distribution of the serial correlation coefficient when $\rho = 0$ has been previously obtained.[2] The purpose of this note is to derive the distribution of the serial correlation coefficient, using the circular definition, when $\rho \neq 0$.

Let us assume that the random variables $x_1, \cdots, x_N$ have a joint normal distribution[3] $p(x_1, \cdots, x_N \mid A, B, \mu)$ where

$$\log p(x_1, \cdots, x_N \mid A, B, \mu) = \log K_1 - \tfrac{1}{2}\left[A \sum_i (x_i - \mu)^2 + 2B \sum_i (x_i - \mu)(x_{i+L} - \mu)\right]$$

the term in the bracket is positive definite, K_1 is independent of the x_i and if $i + L > N$ then $x_{i+L} = x_{i+L-N}$. It is then clear that $\bar{x}$, V_N, and ${}_LC_N$, where $\bar{x}$ is the arithmetic mean, $V_N = \sum_i (x_i - \bar{x})^2$ and

$${}_LC_N = \sum_i (x_i - \bar{x})(x_{i+L} - \bar{x})$$

are sufficient statistics with respect to the estimation of μ, A, and B.

Let $V_N\, {}_LR_N = {}_LC_N$ define ${}_LR_N$, the serial correlation coefficient. Then if

[1] Presented at a meeting of the Cowles Commission for Economic Research in Chicago, January 31, 1945.

[2] See R. L. Anderson, "*Distribution of the serial correlation coefficient*", pp. 1–13 and T. Koopmans, "Serial correlation and quadratic forms in normal variables", pp. 14–33, *Annals of Math. Stat.*, Vol. XIII, No. 1, March, 1942.

[3] The expression $p(\xi_1, \cdots, \xi_m \mid \theta_1, \cdots, \theta_Q)$ means the probability density or the distribution of the random variables $\xi_1, \cdots, \xi_m$ for the given values of the parameters $\theta_1, \cdots, \theta_Q$. When used as an index of summation or multiplication, the letter i will assume all values from 1 through N.

$A = 1, B = 0$ Anderson has shown[4] that, if N is odd, the joint distribution of ${}_1R_N$ and V_N is given by

$$(1)\quad D(R_N, V_N) = KV_N^{\frac{1}{2}(N-3)} e^{-\frac{1}{2}V_N} \sum_{i=1}^{m} (\lambda_i - R_N)^{\frac{1}{2}(N-5)}/\alpha_i, \quad \text{for } \lambda_{m+1} \leq R_N \leq \lambda_m$$

where

$$R_N = {}_1R_N, \qquad \lambda_k = \cos\frac{2\pi k}{N}, \alpha_i = \prod_{j=1}^{\frac{1}{2}(N-1)} (\lambda_i - \lambda_j); \qquad \text{for all } j \neq i$$

and $K^{-1} = 2^{\frac{1}{2}(N-1)}\,\Gamma[\frac{1}{2}(N-3)]$; while if N is even, the same formula holds except that

$$\alpha_i = \prod_{j=1}^{\frac{1}{2}(N-2)} (\lambda_i - \lambda_j)\sqrt{(\lambda_i + 1)}, \qquad \text{for all } j \neq i.$$

We now extend Anderson's distributions to the case where it is not assumed that $A = 1$ and $B = 0$.

As a means of extending[5] Anderson's distribution let us recall that if $x_1, \cdots, x_N$ have a distribution $p(x_1, \cdots, x_N \mid \theta_1, \cdots, \theta_g)$ depending on several parameters $\theta_1, \cdots, \theta_g$, and if $z_1, \cdots, z_k$ are a sufficient set of statistics with respect to $\theta_1, \cdots, \theta_g$, i.e.

$$p(x\ , \cdots, x_N \mid \theta_1, \cdots, \theta_g) = h(z_1, \cdots, z_k \mid \theta_1, \cdots, \theta_g)m(x_1, \cdots, x_N)$$

where $m(x_1, \cdots, x_N)$ is independent of $\theta_1, \cdots, \theta_g$, then if the distribution of $z_1, \cdots, z_k$ is found, assuming $\theta_1, \cdots, \theta_g$ have specific values $\theta_1^0, \cdots, \theta_g^0$, then it follows that

$$p(z_1, \cdots, z_k \mid \theta_1, \cdots, \theta_g) = p(z_1, \cdots, z_k \mid \theta_1^0, \cdots \theta_g^0)\frac{h(z_1, \cdots, z_k \mid \theta_1, \cdots, \theta_g)}{h(z_1, \cdots, z_k \mid \theta_1^0, \cdots, \theta_g^0)}.$$

We may call Anderson's distribution given in (1), $p(R_N, V_N \mid 1, 0)$, i.e.

$$p(R_N, V_N \mid 1, 0) \quad = \quad D(R_N, V_N)$$

Furthermore, $\bar{x}$ is distributed independently of R_N and V_N for all values of A and B and hence by a simple transformation,[6] we can apply the above theorem.

[4] Anderson loc. cit. p. 3 and p. 5. Although the remainder of the note deals only with the case where $L = 1$ the procedure is general and may be easily carried through for other lags.

[5] See W. G. Madow Contributions to the "Theory of multivariate statistical analysis", *Trans. of the Amer. Math. Soc.*, Vol. 44, No. 3, November 1938, p. 461.

[6] For a proof that an orthogonal transformation of the variable $x_i - \mu$ exists such that V_N and ${}_LC_N$ are simultaneously reduced to canonical forms involving the same $N - 1$ of the variables of the transformation, and $\sqrt{N}\ (\bar{x} - \mu)$ is the Nth variable of the transformation, see J. von Neumann, "Distribution of the ratio of the mean square successive difference to the variance, *Annals of Math. Stat.*, Vol. XII, No. 4, December 1941, pp. 368, 369. The proof there is given for V_N and $\sum_i (x_i - x_{i+1})^2$ but is easily extended to this case.

Then it is easy to show that $N(\bar{x} - \mu)$ is independently distributed of V_N, and ${}_LC_N$ and has distribution $\log p[\sqrt{N}(\bar{x} - \mu) \mid A, B] = \log K_2 - \frac{1}{2}[A + 2B]N(\bar{x} - \mu)_2$ where $K_2 = (2\pi)^{-\frac{1}{2}}$ $\cdot(A + 2B)^{\frac{1}{2}}$ and $K_1'K_2 = K_1$.

Then

$$p(R_N, V_N \mid A, B) = p(R_N, V_N \mid 1, 0)\Omega$$

where

$$\Omega = \frac{K_1' e^{-\frac{1}{2}(AV_N + 2BR_NV_N)}}{(2\pi)^{-\frac{1}{2}N} e^{-\frac{1}{2}V_N}}.$$

Hence it follows that,

$$p(R_N, V_N \mid A, B) = KK_1'(2\pi)^{\frac{1}{2}N} V_N^{\frac{1}{2}(N-3)} e^{-\frac{1}{2}V_N(A+2BR_N)} \sum_{i=1}^{m} (\lambda_i - R_N)^{\frac{1}{2}(N-5)}/\alpha_i,$$

for $\lambda_{m+1} \leq R_N \leq \lambda_m$, where the α_i have different values according to whether N is odd or even. In order to evaluate $p(R_N \mid A, B)$ we then need only integrate out V_N. Now

$$\int_0^\infty V_N^{\frac{1}{2}(N-3)} e^{-\frac{1}{2}V_N(A+2BR_N)} \, dV_N = \Gamma[\tfrac{1}{2}(N-1)](A/2 + BR_N)^{-\frac{1}{2}(N-1)}.$$

Hence

$$p(R_N \mid A, B) = KK_1'(2\pi)^{\frac{1}{2}N} \Gamma[\tfrac{1}{2}(N-1)](A/2 + BR_N)^{-\frac{1}{2}(N-1)} \sum_{i=1}^{m} (\lambda_i - R_N)^{\frac{1}{2}(N-5)}/\alpha_i.$$

The parameters K_1', A and B depend on the different types of assumptions that may be made. In general

$$K_1 = (2\pi)^{-\frac{1}{2}N} \Delta^{1/2}$$

where Δ is a circulant $(a_1, \cdots, a_N)$ such that

$$a_1 = A, \qquad a_{1+L} = B, \qquad a_{1+(N-L)} = B, \quad a_i = 0 \text{ otherwise},$$

and hence

$$\Delta = \prod_i \left(A + B\cos\frac{2\pi iL}{N}\right) = \prod (A + B\lambda_i).$$

Then, one assumption is

$$A = \frac{1}{\sigma^2}, \qquad B = -\rho/\sigma^2$$

where ρ is the "true" serial correlation coefficient. Other assumptions are possible.[7] However, these vary with the problem under consideration and may be left for further examination.

---

[7] One possible alternative definition is given by W. J. Dixon, "Further contributions to the problem of serial correlation", *Annals of Math. Stat.*, Vol. XV, No. 2, June 1944, p. 120, equation (2.1).