DISTRIBUTION OF THE SERIAL CORRELATION COEFFICIENT

By R. L. Anderson

North Carolina State College

1. **Introduction.** The problem of serial correlation was brought to the attention of statisticians by Yule in 1921 [9]. Both Yule and Bartlett [2] have shown that the ordinary tests of significance are invalidated if successive observations are not independent of one another. The serial correlation coefficient has been introduced as a measure of the relationship between successive values of a variable ordered in time or space. Interest in the serial correlation problem was stimulated further by the new concepts of time series analysis discussed by Wold [8].

We shall define the serial correlation coefficient for lag L and N observations to be

$$_LR_N = \frac{_LC_N}{V_N} = \frac{X_1X_{L+1} + X_2X_{L+2} + \cdots + X_NX_L - (\Sigma X_i)^2/N}{\Sigma X_i^2 - (\Sigma X_i)^2/N},$$

where C and V are the covariance and variance respectively and the X's are considered to be independently normally distributed about the same mean with unit variance.[1] If the population variance were known a priori, the variates could be transformed so that they would have unit variance; under such an unusual circumstance, the only distribution required would be that of the serial covariance. Tintner has given a test of significance for the serial covariance [6] and for the correlation coefficient [7] by using a method of selected items. The author has presented the distribution of the serial covariance and of the serial correlation coefficient not corrected for the mean in a recent doctoral thesis [1]. The distributions of $_LR_N$ not corrected for the mean will be mentioned in the sections which follow.

2. **Small sample distributions for lag 1.** W. G. Cochran has suggested that we use a result given in his article on quadratic forms to derive the distributions of the serial correlation coefficient for small samples [3]. If $X_1, X_2, \cdots, X_N$ are independently normally distributed with variance 1 and mean 0, then

> "Every quadratic form $\Sigma a_{ij}X_iX_j$ is distributed like $\sum_{k=1}^{r} \lambda_k u_k$, where r is the rank of the matrix, A, of the quadratic form, the u's are independently distributed as χ^2, each with 1 d.f., and the λ's are the non-zero latent roots of the characteristic equation of A" [3, p. 179].

If each λ_i appears k_i times as a latent root, u_i will be distributed as χ^2 with k_i degrees of freedom.

[1] This circular definition of the serial correlation coefficient was suggested by H. Hotelling.

If we set $L = 1$ in the above definition of the serial covariance, we note that the characteristic equation of ${}_1C_N$ is

$$ {}_1F_N = \begin{vmatrix} a_1 & a_2 & a_3 & \cdots & a_N \\ a_N & a_1 & a_2 & \cdots & a_{N-1} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ a_2 & a_3 & a_4 & \cdots & a_1 \end{vmatrix} = 0, $$

where $a_1 = -(\lambda + 1/N)$, $a_2 = a_N = (N - 2)/2N$, and all other a's $= -1/N$. The determinant can be evaluated by the method of circulants. We find that ${}_1F_N(\lambda) = \prod_{K=1}^{N} \left\{ \sum_{i=1}^{N} a_i \omega_k^{i-1} \right\}$, where ω_k is the kth root of unity. Hence,

$$ {}_1F_N = \prod_{K=1}^{N} \left\{ -\left(\lambda_k + \frac{1}{N}\right) + \frac{N-2}{2N}(\omega_k + \omega_k^{-1}) - \frac{1}{N} \sum_{i=3}^{N-1} \omega_k^{i-1} \right\}. $$

Since

$$ \sum_{i=3}^{N-1} \omega_k^{i-1} = \begin{cases} -(\omega_k + 1 + \omega_k^{-1}), & \text{for } k \neq N \\ (N-3), & \text{for } k = N \end{cases} $$

$$ {}_1F_N = \prod_{K=1}^{N-1} \{-\lambda_k + (\omega_k + \omega_k^{-1})/2\} = \prod_{K=1}^{N-1} \left\{ -\lambda_k + \cos \frac{2\pi k}{N} \right\} = 0. $$

Hence $\lambda_k = \cos \dfrac{2\pi k}{N}$, $(k = 1, 2, \cdots, N-1)$, and

$$ {}_1C_N = \begin{cases} \sum_{K=1}^{\frac{1}{2}(N-1)} \lambda_k u_k, & \text{for } N \text{ odd}, \\ \sum_{K=1}^{\frac{1}{2}(N-2)} \lambda_k u_k - u, & \text{for } N \text{ even}, \end{cases} $$

where u_k is distributed as χ^2 with 2 d.f. and u with 1 d.f. At the same time, we note that $V_N = \Sigma(X_i - \bar{X})^2$ is distributed as χ^2 with $N - 1$ d.f.

The general procedure in deriving the distribution of ${}_1R_N$ is as follows: We determine the joint density function of the u's which form the distributions of ${}_1C_N (= {}_1R_N \cdot V_N)$ and V_N. The u's are integrated out, leaving the joint density function of ${}_1R_N$ and V_N. The distribution of ${}_1R_N$ is obtained by integrating with respect to V_N from 0 to ∞. As examples, derivations of the distributions of ${}_1R_6$ and ${}_1R_7$ have been included. In order to simplify the results, the first subscripts have been dropped from ${}_1R_N$.

Distribution of R_6. $R_6V_6 = \lambda_1 u_1 + \lambda_2 u_2 - u$ and $V_6 = u_1 + u_2 + u$, where u_1 and u_2 are distributed as χ^2 with 2 d.f. and u with 1 d.f. and $\lambda_1 = \frac{1}{2}$ and $\lambda_2 = -\frac{1}{2}$. Hence the density function of the u's is

$$ D(u_1, u_2, u) = (4\sqrt{2\pi})^{-1} u^{-\frac{1}{2}} e^{-\frac{1}{2}V_6}. $$

Since $u_1 = [V_6(R_6 - \lambda_2) + u(1 + \lambda_2)]/(\lambda_1 - \lambda_2)$ and

$$u_2 = [V_6(\lambda_1 - R_6) - u(1 + \lambda_1)]/(\lambda_1 - \lambda_2),$$

u must vary between 0 and $V_6(\lambda_1 - R_6)/(1 + \lambda_1)$ for $\lambda_2 \leq R_6 \leq \lambda_1$ and between $V_6(\lambda_2 - R_6)/(1 + \lambda_2)$ and $V_6(\lambda_1 - R_6)/(1 + \lambda_1)$ for $-1 \leq R_6 \leq \lambda_2$. After integrating with respect to u between these limits and then with respect to V_6 from 0 to ∞, we obtained the following density function for R_6 :

$$D(R_6) = \frac{3}{2}\begin{cases} \dfrac{\sqrt{(\lambda_1 - R_6)}}{\sqrt{(1+\lambda_1)}\,(\lambda_1 - \lambda_2)}, & \text{for } \lambda_2 \leq R_6 \leq \lambda_1 \\ \dfrac{\sqrt{(\lambda_1 - R_6)}}{\sqrt{(1+\lambda_1)}\,(\lambda_1 - \lambda_2)} + \dfrac{\sqrt{(\lambda_2 - R_6)}}{\sqrt{(1+\lambda_2)}\,(\lambda_2 - \lambda_1)}, & \text{for } -1 \leq R_6 \leq \lambda_2. \end{cases}$$

The cumulative probability function has the same general form:

$$P(R_6 > R') = \begin{cases} \dfrac{(\lambda_1 - R')^{\frac{3}{2}}}{\sqrt{(1+\lambda_1)}\,(\lambda_1 - \lambda_2)} + \dfrac{(\lambda_2 - R')^{\frac{3}{2}}}{\sqrt{(1+\lambda_2)}\,(\lambda_2 - \lambda_1)} & \text{for } -1 \leq R' \leq \lambda_2 \\ \dfrac{(\lambda_1 - R')^{\frac{3}{2}}}{\sqrt{(1+\lambda_1)}\,(\lambda_1 - \lambda_2)} & \text{for } \lambda_2 \leq R' \leq \lambda_1 \end{cases}$$

Distribution of R_7. $R_7V_7 = \lambda_1u_1 + \lambda_2u_2 + \lambda_3u_3$ and $V_7 = u_1 + u_2 + u_3$, where each u is distributed as χ^2 with 2 d.f. Hence,

$$u_1 = \frac{V_7(R_7 - \lambda_2) + u_3(\lambda_2 - \lambda_3)}{(\lambda_1 - \lambda_2)} \quad \text{and} \quad u_2 = \frac{V_7(\lambda_1 - R_7) - u_3(\lambda_1 - \lambda_3)}{(\lambda_1 - \lambda_2)}.$$

For $\lambda_2 \leq R_7 \leq \lambda_1$, $0 \leq u_3 \leq V_7(\lambda_1 - R_7)/(\lambda_1 - \lambda_3)$; for $\lambda_3 \leq R_7 \leq \lambda_2$, $V_7(\lambda_2 - R_7)/(\lambda_2 - \lambda_3) \leq u_3 \leq V_7(\lambda_1 - R_7)/(\lambda_1 - \lambda_3)$. Using these limits, we derived the following density function for R_7 :

$$D(R_7) = 2 \cdot \begin{cases} \dfrac{(\lambda_1 - R_7)}{(\lambda_1 - \lambda_2)\,(\lambda_1 - \lambda_3)} + \dfrac{(\lambda_2 - R_7)}{(\lambda_2 - \lambda_1)\,(\lambda_2 - \lambda_3)} & \text{for } \lambda_3 \leq R_7 \leq \lambda_2 \\ \dfrac{(\lambda_1 - R_7)}{(\lambda_1 - \lambda_2)\,(\lambda_1 - \lambda_3)} & \text{for } \lambda_2 \leq R_7 \leq \lambda_1. \end{cases}$$

The cumulative probability function is similar, except that the coefficient 2 cancels and the exponent of each numerator is raised by one.

General formulas for N odd. It appears that the density function for R_N and V_N for N odd is

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

where $\alpha_i = \prod_{j=1}^{\frac{1}{2}(N-1)}{}' (\lambda_i - \lambda_j)$ for $j \neq i$ and $1/K = 2^{\frac{1}{2}(N-1)}\Gamma[\frac{1}{2}(N - 3)]$. This

[2] Note that we are omitting the lag subscript from ${}_1R_N$.

formula holds for $N = 5$ and 7; we will show that it holds for $N + 2$, assuming it true for N. If we set $k = \frac{1}{2}(N + 1)$, $R_{N+2}V_{N+2} = R_N V_N + \lambda_k u_k$ and $V_{N+2} = V_N + u_k$; hence,

$$R_N = \frac{(R_{N+2}V_{N+2} - \lambda_k u_k)}{V_{N+2} - u_k} \quad \text{and} \quad V_N = V_{N+2} - u_k.$$

If we make the substitution $u_k = u_k' V_{N+2}$, the density function for u_k', V_{N+2}, and R_{N+2} is

$$\tfrac{1}{2}KV_{N+2}^{\frac{1}{2}(N-1)} e^{-\frac{1}{2}V_{N+2}} \sum_{i=1}^{m} [(\lambda_i - R_{N+2}) - u_k'(\lambda_i - \lambda_k)]^{\frac{1}{2}(N-5)}/\alpha_i.$$

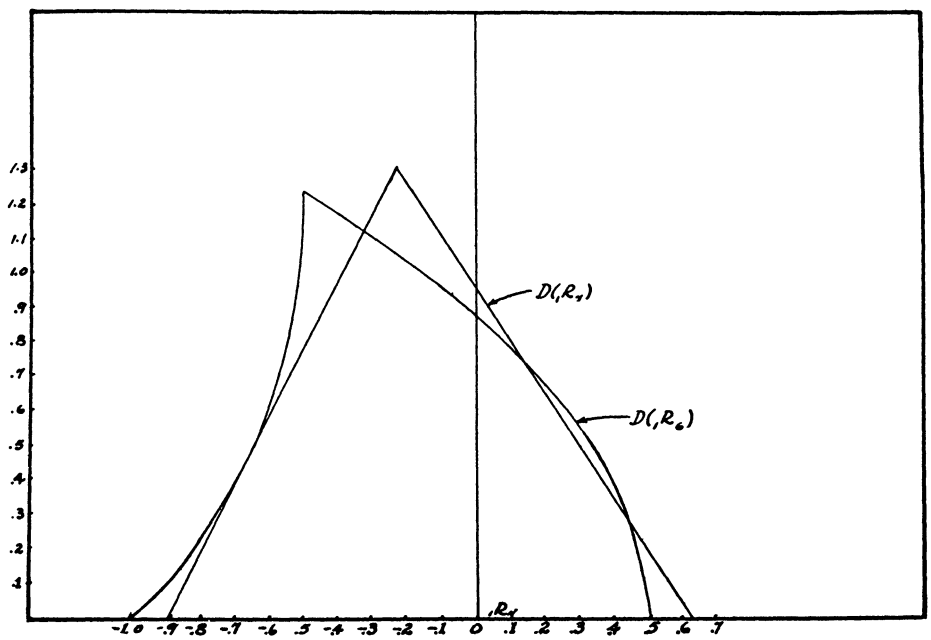

FIG. 1

In order to obtain the distribution of V_{N+2} and R_{N+2}, we must integrate out u_k'. The limits of integration differ for different values of m. We note that

$$u_k' = (R_N - R_{N+2})/(R_N - \lambda_k),$$

except that $u_k' \equiv 0$ when $\lambda_k < R_N \leq \lambda_{m+1}$, since $\lambda_{m+1} \leq R_{N+2} \leq \lambda_m$ and u_k' can not be negative. For $R_{N+2} > \lambda_k$, $u_k' < 1$; hence, if R_N is replaced by a larger (smaller) quantity, u_k' will be larger (smaller).

For $m = 1(\lambda_2 \leq R_{N+2} \leq \lambda_1)$, we need to consider only that region for which $\lambda_2 \leq R_N \leq \lambda_1$. In this region, $0 \leq u_k' \leq (\lambda_1 - R_{N+2})/(\lambda_1 - \lambda_k)$ and the density function of R_{N+2} and V_{N+2} is

$$\phi(V_{N+2})(\lambda_1 - R_{N+2})^{\frac{1}{2}(N-3)}/\alpha_1',$$

where $\phi(V_{N+2}) = V_{N+2}^{\frac{1}{2}(N-1)} e^{-\frac{1}{2}V_{N+2}}/2^{\frac{1}{2}(N+1)} \cdot \Gamma[\frac{1}{2}(N-1)]$ and $\alpha_1' = \prod_{j=2}^{K} (\lambda_1 - \lambda_j)$.

For $m = 2(\lambda_3 \leq R_{N+2} \leq \lambda_2)$, we must consider two regions in the R_N plane. When $\lambda_2 \leq R_N \leq \lambda_1$,

$$\frac{\lambda_2 - R_{N+2}}{\lambda_2 - \lambda_k} \leq u_k' \leq \frac{\lambda_1 - R_{N+2}}{\lambda_1 - \lambda_k},$$

and when $\lambda_3 \leq R_N \leq \lambda_2$, $0 \leq u_k' \leq (\lambda_2 - R_{N+2})/(\lambda_2 - \lambda_k)$. If we combine the density functions for these two regions, we find that

$$D(R_{N+2}, V_{N+2}) = \phi(V_{N+2}) \sum_{i=1}^{2} (\lambda_i - R_{N+2})^{\frac{1}{2}(N-3)}/\alpha_i' \qquad \text{for } \lambda_3 \leq R_{N+2} \leq \lambda_2.$$

Similar results can be obtained for the other regions.

Finally we conclude that for N odd,

$$D({}_1R_N) = \tfrac{1}{2}(N-3) \sum_{i=1}^{m} (\lambda_i - {}_1R_N)^{\frac{1}{2}(N-5)}/\alpha_i \qquad \text{for } \lambda_{m+1} \leq {}_1R_N \leq \lambda_m$$

and

$$P({}_1R_N > R') = \sum_{i=1}^{m} (\lambda_i - R')^{\frac{1}{2}(N-3)}/\alpha_i \qquad \text{for } \lambda_{m+1} \leq R' \leq \lambda_m,$$

where $\alpha_i = \prod_{j=1}^{\frac{1}{2}(N-1)}{}' (\lambda_i - \lambda_j)$, $i \neq j$. The general density function for N odd and ${}_1R_N$ not corrected for the sample mean is [1]

$$D({}_1R_N) = \tfrac{1}{2}(N-2) \sum_{i=m}^{\frac{1}{2}(N-1)} ({}_1R_N - \lambda_i)^{\frac{1}{2}(N-4)}/\alpha_i \quad \text{for} \quad \lambda_m \leq {}_1R_N \leq \lambda_{m-1},$$

where $\alpha_i = \prod_{j=1}^{\frac{1}{2}(N-1)}{}' (\lambda_j - \lambda_i)\sqrt{(1-\lambda_i)}$, $i \neq j$.

General formulas for N even. Using the same method as above, we can show that the same formulas hold for N even and ${}_1R_N$ corrected for the mean except that in this case $\alpha_i = \prod_{j=1}^{\frac{1}{2}(N-2)}{}' (\lambda_i - \lambda_j)\sqrt{(\lambda_i + 1)}$, $j \neq i$. No general formulas were derived for N even and ${}_1R_N$ not corrected for the mean.

3. **Large sample distributions for lag 1.** The simultaneous density function of C and V, where we will drop the subscripts for convenience, is

$$D(C, V) = (2\pi)^{-2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(s, t) e^{-sC - tV}\, ds\, dt.$$

$$\phi(s, t) = K \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\theta}\, dX_1\, dX_2 \cdots dX_N,$$

where $\theta = \{\Sigma X_i^2 - 2t[\Sigma(X_i - \bar{X})^2] - 2s[X_1X_2 + \cdots + X_NX_1 - (\Sigma X_i)^2/N]\}$ and s and t are pure imaginaries.

$\phi(s, t) = \Delta^{-\frac{1}{2}}$, where Δ is the determinant of the quadratic form θ. This determinant was evaluated by the method of circulants; we found that $\Delta = \prod_{k=1}^{N-1} \{1 - 2(t + s\lambda_k)\}$, where $\lambda_k = \cos 2\pi k/N$.

Set $K = \log \phi(s, t) = \Sigma \kappa_{ij} \frac{s^i t^j}{i!j!}$. If K is expanded in series, we find that $\kappa_{ij} = m!2^m \sum_{k=1}^{N-1} \lambda_k^i$, where $m = (i + j - 1)$. For $N > i$, we might indicate these summations: $\Sigma\lambda_k = -1$, $\Sigma\lambda_k^2 = \frac{1}{2}(N - 2)$, $\Sigma\lambda_k^3 = -1$, $\Sigma\lambda_k^4 = \frac{1}{8}(3N - 8)$ and $\Sigma\lambda_k^5 = -1$. Hence $\kappa_{10} = E(C) = -1$, $\kappa_{01} = E(V) = (N - 1)$, $\kappa_{20} = \sigma_c^2 = (N - 2)$, $\kappa_{02} = \sigma_v^2 = 2(N - 1)$, $\kappa_{11} = \rho\sigma_c\sigma_v = -2$, $\kappa_{30} = -8$, $\kappa_{03} = 8(N - 1)$, $\kappa_{21} = 4(N - 2)$, $\kappa_{12} = -8$, etc.

If we let $C' = C + 1$ and $V' = V - (N - 1)$, all of these semi-invariants will remain unchanged except that $\kappa_{10} = \kappa_{01} = 0$. Since $R = C/V$,

$$\left(R + \frac{1}{N - 1}\right) = \frac{C'(N - 1) + V'}{[V' + (N - 1)](N - 1)}$$

$$= \frac{C'(N - 1) + V'}{(N - 1)^2} \left\{\sum_{p=0}^{\infty} (-1)^p \left(\frac{V'}{N - 1}\right)^p\right\}.$$

If we neglect terms of order less than $1/N$, $E(R) = -1/(N - 1)$, $E(R - \bar{R})^2 = \frac{(N - 2)}{(N - 1)^2}$, and $E(R - \bar{R})^k = 0$ for $k > 2$. For $N < 75$, a more exact approximation may be desired.

If the above approximation is used, ${}_1R_N$ is normally distributed with mean $-1/(N - 1)$ and variance $(N - 2)/(N - 1)^2$. The single-tail significance points can be found by substituting in the formulas

$${}_1R_N(.05) = \frac{-1 \pm 1.645\sqrt{(N - 2)}}{N - 1} \quad \text{or} \quad {}_1R_N(.01) = \frac{-1 \pm 2.326\sqrt{(N - 2)}}{N - 1}$$

Refer to Fig. 2 for a comparison of the exact distribution and the normal approximation for $N = 15$. I have included the graphs of the exact distributions for $N = 6$ and 7 in Fig. 1. We might note a few comparisons between the approximate significance points and the exact ones:

N	Positive tail				Negative tail			
	5%		1%		5%		1%	
	Exact	Approx.	Exact	Approx.	Exact	Approx.	Exact	Approx.
45	0.218	0.223	0.314	0.324	−0.262	−0.268	−0.356	−0.369
75	0.173	0.176	0.250	0.255	−0.199	−0.203	−0.276	−0.282

For ${}_1R_N$ not corrected for the mean, it was found that $y = \sqrt{\frac{N\,{}_1R_N^2}{1 + 2\,{}_1R_N^2}}$ was asymptotically normally distributed with mean 0 and variance 1 [1].

4. **Significance points of ${}_1R_N$.** An example of the methods used in tabulating these significance points has been presented in the author's doctoral thesis [1]. The significance points for the values of N enclosed in parentheses have been obtained by graphical interpolation. Note that N is the number of observations (see Table I).

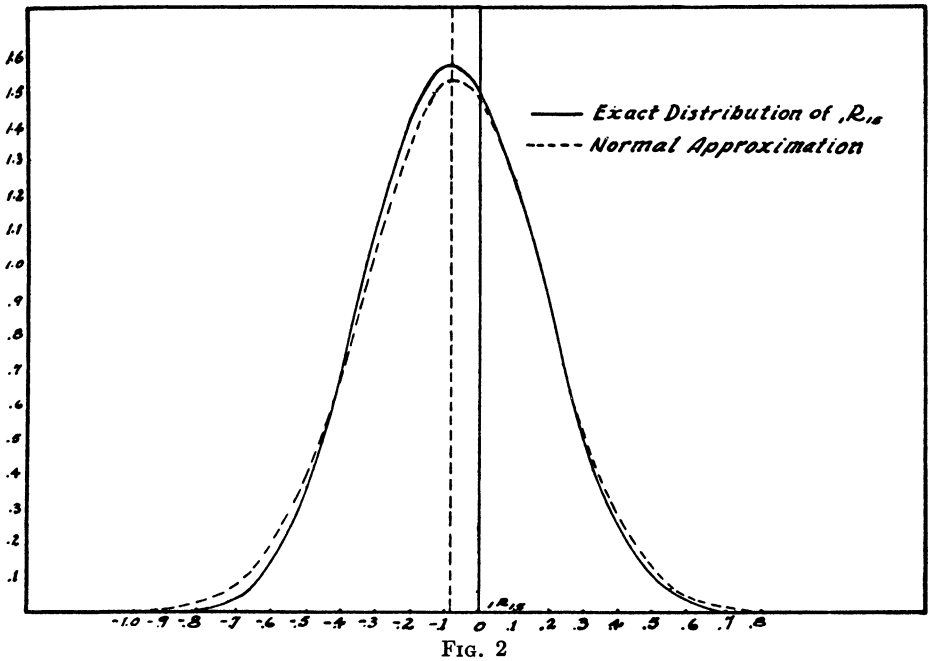

FIG. 2

5. **Distributions for general lag, L.** (a) *Introduction.* For a general lag, L, the constants in the characteristic equation for the covariance ${}_LC_N$ are $a_1 = -(\lambda + 1/N)$, $a_{L+1} = a_{N-L+1} = (N - 2)/2N$ and all other a's $= -1/N$. Hence the characteristic equation is

$$ {}_LF_N = \prod_{k=1}^{N-1} [\lambda_k - \cos\cdot(2\pi Lk/N)] = 0. $$

Certain important generalizations concerning ${}_LF_N$ may be set down:

1. When L is not a factor of N or has no common factor with N, ${}_LF_N = {}_1F_N$.
2. When L and N have a common factor, α, ${}_LF_N = ({}_1F_{N/\alpha})^{\alpha}(\lambda - 1)^{\alpha-1}$.

2a. If $\alpha = L$, ${}_LF_N = ({}_1F_p)^L(\lambda - 1)^{L-1}$, where $p = N/L$.

The proof of the first statement was suggested by Cochran. Since $\cos(\alpha + 2a\pi) = \cos\alpha$, where a is any integer we must prove that the series of numbers

$$L, 2L, \cdots, (N-1)L,$$

when reduced modulus N can be arranged to form the series

$$1, 2, \cdots, (N-1).$$

This proof can be found in most books on the theory of numbers; e.g. [4]. Hence we conclude that each term of the sequence $\{\cos(2\pi Lk/N)\}$ reduces uniquely

TABLE I

N	Positive tail		Negative tail	
	5%	1%	5%	1%
5	0.253	0.297	−0.753	−0.798
6	0.345	0.447	0.708	0.863
7	0.370	0.510	0.674	0.799
8	0.371	0.531	0.625	0.764
9	0.366	0.533	0.593	0.737
10	0.360	0.525	0.564	0.705
11	0.353	0.515	0.539	0.679
12	0.348	0.505	0.516	0.655
13	0.341	0.495	0.497	0.634
14	0.335	0.485	0.479	0.615
15	0.328	0.475	0.462	0.597
20	0.299	0.432	0.399	0.524
25	0.276	0.398	0.356	0.473
30	0.257	0.370	0.325	0.433
(35)	0.242	0.347	0.300	0.401
(40)	0.229	0.329	0.279	0.376
45	0.218	0.314	0.262	0.356
(50)	0.208	0.301	0.248	0.339
(55)	0.199	0.289	0.236	0.324
(60)	0.191	0.278	0.225	0.310
(65)	0.184	0.268	0.216	0.298
(70)	0.178	0.259	0.207	0.287
75	0.173	0.250	−0.199	−0.276

to one of the sequence $\{\cos(2\pi k/N)\}$ for $k = 1, 2, \cdots, (N-1)$, when L/N is a prime fraction.

If L and N have a common factor, α, $L = q\alpha$ and $N = p\alpha$, where p and q are integers prime to one another. Hence,

$$
\begin{aligned}
{}_LF_N &= \prod_{k=1}^{p\alpha-1}\left\{\lambda_k - \cos\frac{2\pi qk}{p}\right\} = \prod_{k=1}^{p-1}\left(\lambda_k - \cos\frac{2\pi k}{p}\right)^{\alpha}(\lambda - \cos 2\pi)^{\alpha-1} \\
&= ({}_1F_p)^{\alpha}(\lambda - 1)^{\alpha-1} = 0.
\end{aligned}
$$

If $\alpha = L$, ${}_LF_N = ({}_1F_p)^{L}(\lambda - 1)^{L-1}$, where $p = N/L$.

When these results are applied to the large sample distribution of ${}_LR_N$, we find that it is independent of L. For the more important case in which $p = N/L$, the semi-invariants κ_{ij} for C and V are exactly the same for all L with a given N. We see that

$$K_L = -\tfrac{1}{2}L \sum_{k=1}^{p-1} \log \{1 - 2(t + s\lambda_k)\} - \tfrac{1}{2}(L - 1) \log \{1 - 2(t + s)\},$$

where $\lambda_k = \cos(2\pi k/p)$. Hence, $\kappa_{ij} = m!2^m \left\{ \frac{N}{p} \left(\sum_{k=1}^{p-1} \lambda_k^i + 1 \right) - 1 \right\}$. But $\sum_{k=1}^{p-1} \lambda_k^i + 1$ is always 0 or a multiple of p when $p > i$; therefore, the p's cancel and κ_{ij} is the same for all p or for all L, since $L = N/p$. When $p \leq i$, the κ_{ij}'s will not be equal for all p. For example $\kappa_{20} = 2(N - 1)$ for $p = 2$ and $\kappa_{30} = 2(N - 4)$ for $p = 3$.

(b) *Distributions of* ${}_LR_N$ *when* $N/L = p$. These results indicate that the distributions of the serial correlation coefficients for which the number of observations is divisible by the lag, so that $N/L = p$, would include the distributions of all the serial correlation coefficients regardless of the values of N and L. We will designate any lag L as the primary lag for a given N if $N/L = p$, an integer. For example, ${}_2R_6$ and ${}_4R_6$ have the same density function, but we will derive only the density function for lag 2, which we will call the primary lag. The case of $p = 1$ is trivial, since it involves correlating a series with itself. To date, we have derived the exact density functions for $p = 2$ and $p = 3$ and the required integrals for $p = 4$. The significance points have been tabulated in Table II. For simplicity of notation, we will set ${}_LR_N = {}_LR_p$ and $V_N = V$.

Case $p = 2(N = 2L)$. ${}_LR_2V = -u_1 + u_2$ and $V = u_1 + u_2$, where u_1 is distributed as χ^2 with L d.f. and u_2 as χ^2 with $L - 1$ d.f. Hence,

$$D_L(u_1, u_2) = K(u_1)^{\frac{1}{2}(L-2)}(u_2)^{\frac{1}{2}(L-3)},$$

where $1/K = 2^{L-\frac{1}{2}}\Gamma(\frac{1}{2}L)\Gamma[\frac{1}{2}(L - 1)]e^{V/2}$. After substituting $u_1 = V(1 - {}_LR_2)/2$ and $u_2 = V(1 + {}_LR_2)/2$ and integrating with respect to V from 0 to ∞, we have

$$D({}_LR_2) = \frac{(1 - {}_LR_2)^{\frac{1}{2}(L-2)}(1 + {}_LR_2)^{\frac{1}{2}(L-3)}}{2^{L-\frac{1}{2}}\beta[\frac{1}{2}L, \frac{1}{2}(L - 1)]}.$$

If we set $(1 - {}_LR_2) = 2y$, then the cumulative probability function is

$$P({}_LR_2 > R') = \frac{1}{\beta[\frac{1}{2}L, \frac{1}{2}(L - 1)]} \int_{y=0}^{\frac{1}{2}(1-R')} y^{\frac{1}{2}(L-2)}(1 - y)^{\frac{1}{2}(L-3)}\, dy.$$

Pearson has tabulated the values of these incomplete Beta functions [5]. In his notation, $P = I_x[\frac{1}{2}L, \frac{1}{2}(L - 1)]$, where $x = \frac{1}{2}(1 - R')$. For ${}_LR_2$ not corrected for the mean, $P = I_x(\frac{1}{2}L, \frac{1}{2}L)$ [1].

Case $p = 3(N = 3L)$. ${}_LR_3V = -\frac{1}{2}u_1 + u$ and $V = u_1 + u$, where u_1 is distributed as χ^2 with $2L$ d.f. and u with $L - 1$ d.f. Therefore, $D_L(u_1, u) =$

$Ku_1^{L-1}u^{\frac{1}{2}(L-3)}$, where $1/K = 2^{\frac{1}{2}(3L-1)}\Gamma(L)\Gamma[\frac{1}{2}(L-1)]e^{V/2}$. After substituting $u_1 = 2V(1 - {}_LR_3)/3$ and $u = V(1 + 2{}_LR_3)/3$ and integrating with respect to V from 0 to ∞, we find that

$$D({}_LR_3) = \frac{2^L(1 - {}_LR_3)^{L-1}(1 + 2{}_LR_3)^{\frac{1}{2}(L-3)}}{3^{\frac{1}{2}(L-1)}\beta[L, \frac{1}{2}(L-1)]}, \qquad {}_LR_3 \geq -\tfrac{1}{2}.$$

If we set $x = 2(1 - R')/3$, $P({}_LR_3 > R') = I_x[L, \frac{1}{2}(L-1)]$. For ${}_LR_3$ not corrected for the mean, $P = I_x[L, \frac{1}{2}L]$.

Case $p = 4(N = 4L)$. ${}_LR_4V = -u_2 + u_4$ and $V = u_2 + u_4 + u$, where u_2 is distributed as χ^2 with L d.f., u_4 with $L - 1$ d.f. and u with $2L$ d.f. The density function of the u's is $D_L(u_2, u_4, u) = Ku_2^{\frac{1}{2}(L-2)}u_4^{\frac{1}{2}(L-3)}u^{L-1}e^{-V/2}$, where $1/K = 2^{\frac{1}{2}(4L-1)}\Gamma(\frac{1}{2}L)\Gamma[\frac{1}{2}(L-1)]\Gamma(L)$. Since $u_4 = [V(1 + {}_LR_4) - u]/2$ and $u_2 = [V(1 - {}_LR_4) - u]/2$, $0 \leq u \leq V(1 - {}_LR_4)$ for ${}_LR_4 \geq 0$ and $0 \leq u \leq V(1 + {}_LR_4)$ for ${}_LR_4 \leq 0$. For ${}_LR_4 \geq 0$,

$$D({}_LR_4) = \frac{KVe^{-\frac{1}{2}V}}{2^{\frac{1}{2}(2L-3)}} \int_{u=0}^{V(1-{}_LR_4)} [V(1 + {}_LR_4) - u]^{\frac{1}{2}(L-3)}[V(1 - {}_LR_4) - u]^{\frac{1}{2}(L-2)}\, u^{L-1}\, du.$$

For ${}_LR_4 \leq 0$, $D({}_LR_4)$ is the same except that the upper limit for the integral is $V(1 + {}_LR_4)$. If we make the substitution $y = u/(\text{upper limit})$ in each case and then integrate with respect to V from 0 to ∞, we have these density functions:

$$D({}_LR_4) = k \cdot \begin{cases} (1 + {}_LR_4)^{\frac{1}{2}(3L-3)} \displaystyle\int_{y=0}^{1} y^{L-1}(1-y)^{\frac{1}{2}(L-3)}[(1 - {}_LR_4) - y(1 + {}_LR_4)]^{\frac{1}{2}(L-2)}\, dy, \\ \qquad\qquad \text{for } {}_LR_4 \leq 0, \\ (1 - {}_LR_4)^{\frac{1}{2}(3L-2)} \displaystyle\int_{y=0}^{1} y^{L-1}(1-y)^{\frac{1}{2}(L-2)}[(1 + {}_LR_4) - y(1 - {}_LR_4)]^{\frac{1}{2}(L-3)}\, dy, \\ \qquad\qquad \text{for } {}_LR_4 \geq 0, \end{cases}$$

where $k = \Gamma[\frac{1}{2}(4L-1)]/2^{\frac{1}{2}(2L-3)} \cdot \Gamma(L) \cdot \Gamma(\frac{1}{2}L) \cdot \Gamma[\frac{1}{2}(L-1)]$.

The probability integrals must be evaluated for each L. The cumulative probability functions for $L = 2$ and 3 are:

$$P({}_LR_4 > R') = 1 - \frac{\sqrt{2}}{2} \cdot \begin{cases} (1 + R')^{5/2} - R'^{3/2}(5 + R')/\sqrt{2}, & \text{for } R' \geq 0, \\ (1 + R')^{5/2}, & \text{for } R' \leq 0, \end{cases}$$

$$P({}_LR_4 > R') = \frac{\sqrt{2}}{4} \begin{cases} (1 - R')^{9/2}, & \text{for } R' \geq 0, \\ (1 - R')^{9/2} - (-R'/2)^{5/2}(22R'^2 + 36R' + 126), & \text{for } R' \leq 0. \end{cases}$$

Since the density functions are much simpler for $R' > 0$ when L is odd and for $R' < 0$ when L is even, we have derived only these significance points for $L > 3$ and interpolated for the intermediate points. It was noted that the significance points approach those given in Table I for the first lag. For these comparisons, see Table III below. Note that for $L \geq 7$ the 5% points are almost identical and the 1% points are nearly accurate to two decimal places.

TABLE II

Significance points of ${}_LR_N$ for $p = 2$ and 3[3]

L[3]	p=2 ($N=2L$)				p=3 ($N=3L$)			
	Positive tail		Negative tail		Positive tail		Negative tail	
	5%	1%	5%	1%	5%	1%	5%	1%
2	0.805	0.960	−0.99	−1.00	0.488	0.762	−0.496	−0.50
3	0.729	0.907	0.928	0.994	0.447	0.677	0.474	0.496
4	0.604	0.852	0.848	0.950	0.406	0.610	0.439	0.480
5	0.612	0.802	0.773	0.902	0.373	0.559	0.406	0.461
6	0.571	0.759	0.712	0.856	0.346	0.518	0.377	0.440
7	0.536	0.721	0.662	0.812	0.324	0.485	0.354	0.420
8	0.507	0.688	0.620	0.774	0.306	0.457	0.334	0.402
9	0.483	0.659	0.585	0.739	0.291	0.433	0.316	0.387
10	0.462	0.634	0.554	0.708	0.278	0.413	0.301	0.373
12	0.428	0.590	0.505	0.656	0.256	0.380	0.276	0.347
14	0.399	0.554	0.467	0.612	0.239	0.353	0.256	0.326
16	0.376	0.523	0.436	0.577	0.225	0.332	0.240	0.308
18	0.357	0.498	0.410	0.546	0.213	0.314	0.227	0.293
20	0.340	0.476	0.389	0.520	0.202	0.298	0.215	0.280
25	0.308	0.432	0.347	0.469	0.182	0.268	0.193	0.254
30	0.282	0.398	0.317	0.431	0.167	0.245	0.176	0.234
40	0.247	0.348	0.273	0.374	0.146	0.212	0.153	0.205
50	0.222	0.314	−0.243	−0.335	0.131	0.191	−0.136	−0.184

TABLE III[4]

Significance points for $p = 4$

L	N	Positive tail				Negative tail			
		5%		1%		5%		1%	
		Exact	Table 1	Exact	Table 1	Exact	Table 1	Exact	Table 1
2	8	0.373	0.371	0.618	0.531	−0.653	−0.625	−0.818	−0.764
3	12	0.353	0.348	0.547	0.505	0.528	0.516	0.692	0.655
4	16	0.325*	0.322	0.490*	0.466	0.451	0.447	0.604	0.580
5	20	0.301	0.299	0.451	0.432	0.402*	0.409	0.543*	0.524
6	24	0.281*	0.280	0.419*	0.404	0.365	0.363	0.497	0.482
7	28	0.264	0.264	0.392	0.380	−0.338*	−0.337	−0.460*	−0.448

[3] L is the lag and $p = N/L$.

[4] * indicates interpolated values.

Case $p > 4$. We have not set up any of the density functions for $p > 4$; however, it appears that the significance points given for lag 1 would be accurate enough for the higher lags. The exact significance points for lag 2 have been derived for $p = 5$ and 7. The reader may note the close approximation given by the significance points for lag 1 when $p = 7$. We hope to check the lag 1 approximation for other lags in the near future.

TABLE IV

Some significance points for lag 2

	Positive tail		Negative tail	
	5%	1%	5%	1%
$p = 5$ $(N = 10)$				
Exact	0.342	0.540	−0.417	−0.595
Approx.	0.360	0.525	−0.564	−0.705
$p = 7$ $(N = 14)$				
Exact	0.335	0.482	−0.479	−0.616
Approx.	0.335	0.485	−0.479	−0.615

7. **Summary.** 1. The exact and large sample distributions have been derived for the serial correlation coefficient for lag 1 and the exact significance points tabultaed for N, the number of observations, up to 75; for $N > 75$, the large sample approximations can be used.

2. It has been noted that the distributions for any lag L are the same as those for lag 1 when L and N are prime to each other. In general the distribution of the serial correlation coefficient can be derived for any L and N by using only those distributions for which L is a factor of N. The distributions and significance points have been derived for $N/L = p = 2, 3$ and 4. For $p > 4 (N > 4L)$, the significance points given for lag 1 probably can be used when L is greater than 4 or 5. The accuracy of this approximation has been checked for lag 2.

3. These significance points should be useful in determining the methods of studying a time series, as suggested by Wold, and in the formulation of a better test of the significance of regression coefficients when we know that the observations are correlated in time. In addition, we now have a method of testing our assumptions of independence for any set of data.

REFERENCES

[1] R. L. ANDERSON, *Serial Correlation in the Analysis of Time Series*, unpublished thesis, Library, Iowa State College, Ames, Iowa, 1941.

[2] M. S. BARTLETT, "Some aspects of the time-correlation problem in regard to tests of significance," *Roy. Stat. Soc. Jour.*, Vol. 98 (1935), pp. 536-543.

[3] W. G. COCHRAN, "Distribution of quadratic forms in a normal system with applications to the analysis of covariance," *Camb. Phil. Soc. Proc.*, Vol. 30 (1934), pp. 178–191.
[4] L. E. DICKSON, *Modern Elementary Theory of Numbers*, U. of Chicago Press, 1939.
[5] KARL PEARSON (Editor), *Tables of the Incomplete Beta-Function*, Cambridge U. Press, 1934.
[6] G. TINTNER, "Tests of significance in time series," *Annals of Math. Stat.*, Vol. 10 (1939), p. 141 ff.
[7] G. TINTNER, *The Variate Difference Method*, Principia Press, Bloomington, Indiana, Appendix 5B, 1940.
[8] H. WOLD, *A Study in the Analysis of Stationary Time Series*, Almquist and Wiksells Boktryckeri A. B., Uppsala, 1939.
[9] G. U. YULE, "On the time-correlation problem," *Roy. Stat. Soc. Jour.* Vol. 84 (1921), pp. 496–537.