EXACT TESTS OF SERIAL CORRELATION USING NONCIRCULAR STATISTICS

By G. S. Watson and J. Durbin

University of Cambridge and London School of Economics

1. Summary and introduction. For testing the hypothesis that successive members of a series of observations are independent J. von Neumann [5] (see also B. I. Hart [4]) and R. L. Anderson [1] have proposed test statistics and tabulated their significance points. von Neumann's criterion seems well designed to detect deviations from the null hypothesis which might be encountered in practice but its exact distribution is unknown. On the other hand Anderson's statistic, while it has a known distribution, is based on a circular conception of the population which is rarely plausible in practice.

In the present note certain noncircular statistics are proposed for which exact distributions can be obtained from Anderson's results. The statistics are derived from the usual noncircular statistics by sacrificing a small amount of relevant information. Their application is noted to certain regression problems for which no satisfactory tests are at present available. Finally, some general remarks are made about the choice of best statistics for the problems discussed.

2. Proposed statistics. Consider the ratio

$$r = \frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}}, \tag{1}$$

where $\boldsymbol{x} = \{x_1 x_2 \cdots x_n\}$ is a column vector of independent normal variables with zero means and constant variance and $\boldsymbol{A}$ is a real symmetric matrix. Then the exact distribution of r is at present known only when the characteristic roots of $\boldsymbol{A}$ all have the same even multiplicity except at most two of arbitrary multiplicity. Thus in particular the distribution of r is known if $\boldsymbol{A}$ can be written as

$$\begin{pmatrix} \boldsymbol{B} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \lambda \boldsymbol{I}_p & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{B} \end{pmatrix}, \tag{2}$$

where $\boldsymbol{B}$ is a real symmetric $l \times l$ matrix with distinct roots $\nu_1 > \nu_2 > \cdots > \nu_l$, $\boldsymbol{I}_p$ is the unit matrix of order p, and λ satisfies either $\lambda \leq \nu_l$ or $\lambda \leq \nu_1$. Using the results of R. L. Anderson [1], we give the distribution of r when λ is not equal to ν_l or ν_1. For $\lambda < \nu_l$,

$$P(r > r') = \sum_{i=1}^{s} \frac{(\nu_i - r')^{l+\frac{1}{2}p-1}}{(\nu_i - \lambda)^{\frac{1}{2}p} \prod_j{}' (\nu_i - \nu_j)} \qquad (\nu_{s+1} \leq r' \leq \nu_s), \tag{3}$$

and for $\lambda > \nu_1$,

$$P(r > r') = 1 - \sum_{i=s+1}^{l} \frac{(r' - \nu_i)^{l+\frac{1}{2}p-1}}{(\lambda - \nu_i)^{\frac{1}{2}p} \prod_j{}' (\nu_j - \nu_i)} \qquad (\nu_{s+1} \leq r' \leq \nu_s), \tag{4}$$

where

$$\prod_j{}' (\nu_i - \nu_j) = \prod_{\substack{j=1 \\ j\neq i}}^{l} (\nu_i - \nu_j),$$

$$\prod_j{}' (\nu_j - \nu_i) = \prod_{\substack{j=1 \\ j\neq i}}^{l} (\nu_j - \nu_i).$$

If λ equals ν_l or ν_1, the distribution of r may be obtained from (3) or (4) by a renumbering of the roots. These expressions remain correct for $p = 0$ or

$$A = \begin{pmatrix} B & 0 \\ 0 & B \end{pmatrix}.$$

The probability densities can be derived by differentiation.

For testing the hypothesis of serial independence in a set of observations with a zero (or known) mean we propose the following statistics:

$$c_1 = \frac{x_1x_2 + \cdots + x_{m-1}x_m + x_{m+1}x_{m+2} + \cdots + x_{n-1}x_n}{\sum_1^n x_i^2} \qquad (n = 2m) \tag{5}$$

and

$$c_2 = \frac{x_1x_2 + \cdots + x_{m-1}x_m + x_{m+1}^2 + x_{m+2}x_{m+3} + \cdots + x_{n-1}x_n}{\sum_1^n x_i^2} \qquad (n = 2m + 1)$$

or

$$c_2' = \frac{x_1x_2 + \cdots + x_{m-1}x_m + x_{m+2}x_{m+3} + \cdots + x_{n-1}x_n}{\sum_1^n x_i^2 - x_{m+1}^2} \qquad (n = 2m + 1).$$

It is easily seen that these statistics can be written in the form (1) with A having the form (2). Here

$$B = \tfrac{1}{2}\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & & & \\ \cdot & \cdot & \cdot & & & \\ \cdot & \cdot & \cdot & & & \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix},$$

which has characteristic roots $\nu_i = \cos \dfrac{\pi i}{m+1}$ $(i = 1, \cdots, m)$. For c_2, $p = 1$

and $\lambda = 1$. Thus the distribution of c_1 is given by (3) and (4) with these ν_i's, $l = m$ and $p = 0$. The distribution of c_2 is given by (4) with $l = m$, $p = 1$, and $\lambda = 1$. In c_2', x_{m+1} has merely been omitted altogether so that c_2 has the same form and therefore the same distribution as c_1.

An alternative set of statistics is

$$d_1 = \frac{(x_1 - x_2)^2 + \cdots + (x_{m-1} - x_m)^2 + (x_{m+1} - x_{m+2})^2 + \cdots + (x_{n-1} - x_n)^2}{\sum_1^n x_i^2} \qquad (n = 2m)$$

and

$$d_2 = \frac{(x_1 - x_2)^2 + \cdots + (x_{m-1} - x_m)^2 + (x_{m+2} - x_{m+3})^2 + \cdots + (x_{n-1} - x_n)^2}{\sum_1^n x_i^2} \qquad (n = 2m + 1)$$

or

$$d_2' = \frac{\text{numerator of } d_2}{\sum_1^n x_i^2 - x_{m+1}^2} \qquad (n = 2m + 1).$$

As before these may be thrown into the form (1) with

$$B = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & & & \\ \cdot & \cdot & \cdot & & & \\ \cdot & \cdot & \cdot & & & \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix},$$

which has characteristic roots $\nu_1 = 4 \sin^2 \frac{(m - i)\pi}{2m}$ $(i = 1, \cdots, m)$. Here $\lambda = 0$. Thus the distribution of d_1 and d_2' is given by (3) and (4) with these ν_i's, $l = m$ and $p = 0$, while the distribution of d_2 is given by (3) with $l = m - 1$, $p = 3$, and $\lambda = 0$.

For the case of an unknown mean we propose the statistics

(6)
$$d_3 = \frac{\text{numerator of } d_1}{\sum_1^n (x_i - \bar{x})^2} \qquad (n = 2m)$$

and

$$d_4 = \frac{\text{numerator of } d_2}{\sum_1^n (x_i - \bar{x})^2} \qquad (n = 2m + 1)$$

or d_4' $(n = 2m + 1)$ which is of the same form as d_3 (the central observation in a series of $2m + 1$ being omitted). The distribution of d_3 and d_4' is given by (3) with

$$\nu_i = 4 \sin^2 \frac{(m - i)\pi}{2m} \quad (i = 1, \cdots, m - 1), \quad l = m - 1, \quad p = 1, \text{ and } \lambda = 0.$$

The distribution of d_4 is given by (3) and (4) with

$$\nu_i = 4 \sin^2 \frac{(m - i)\pi}{2m} \qquad (i = 1, \cdots, m) \; l = m \text{ and } p = 0.$$

The distinction between the above statistics, exemplified by c_1, and the related circular statistics, exemplified by

$$c_3 = \frac{x_1x_2 + \cdots + x_{n-1}x_n + x_nx_1}{\sum_1^n x_i^2},$$

is now clear. In each case the numerator quadratic form has been modified from the obvious form to take, namely $\Sigma_2^n x_ix_{i-1}$, to a form giving a statistic with a known distribution. In c_1 this is done by throwing out the relevant term x_mx_{m+1}, whereas in c_3 an extraneous term, x_nx_1, is included.

3. Application to regression problems. Suppose we have the sample corresponding to the regression equation

$$y_t = \beta_1x_{1t} + \cdots + \beta_kx_{kt} + \epsilon_t \qquad (t = 1, \cdots, n),$$

and we wish to test the ϵ_t for serial correlation. Exact tests are at present available only for cases in which the characteristic roots of $\boldsymbol{A}$ in (1) occur with certain multiplicities mentioned in Section 2, and the regression vectors $x_1, \cdots, x_k$ are linear functions of a suitable set of k of the characteristic vectors of $\boldsymbol{A}$. (T. W. Anderson [2].) Such cases are rare in practice. For other cases the general problem has been discussed elsewhere (Durbin and Watson [3]). We suggest here an approach that will give an exact test when the x vectors are the same in different applications, as for example in polynomial regressions and analysis of variance models.

For $n = 2m$ or $n = 2m + 1$ suppose that separate least squares regression analyses are carried out on the first m and the last m observations. We shall confine ourselves to cases in which the regression vectors are the same in the two analyses. We may, for instance, have fitted a parabolic trend separately to each of the two sets of observations. Consider

$$r = \frac{z_1'\boldsymbol{B}z_1 + z_2'\boldsymbol{B}z_2}{z_1'z_1 + z_2'z_2},$$

where z_1 and z_2 are the two sets of residuals from regression, and $\boldsymbol{B}$ is a real symmetric matrix with distinct roots. It has been shown (Durbin and Watson [3]) that r is distributed as

$$\sum_{i=1}^{m-k} \mu_i(\eta_{1i}^2 + \eta_{2i}^2) \Big/ \sum_{i=1}^{m-k} (\eta_{1i}^2 + \eta_{2i}^2),$$

where the η's are independent normal variables with zero means and unit variances and $\mu_1, \cdots, \mu_{m-k}$ are the characteristic roots other than k zeros of the matrix $(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}^{-1})\boldsymbol{X}')\boldsymbol{B}$. $\boldsymbol{X}$ here is the $m \times k$ matrix of independent variables used in the subanalyses. Either of the two forms of $\boldsymbol{B}$ given in Section 2 may be used. If the roots are known the distribution of r can be obtained from (3) or (4).

These results have applications to a number of problems in time series analysis. It is proposed, for example, to calculate the characteristic roots, and hence construct an exact test, for the residuals from a polynomial trend. Other regression models that could be treated in a similar way are one- and two-way classifications and periodic regressions.

We might mention that the fitting of separate regressions to the two halves of the series will often be less artificial than might at first sight appear since it is in any case a common practice to break up time series into two or more parts for the fitting of trends.

4. Powers of the statistics. T. W. Anderson [2] has discussed the Neyman-Pearson theory for testing the hypothesis of serial independence of the error terms of a regression equation. In testing for serial independence against the alternative that the errors follow a stationary (normal) first order autoregressive scheme

$$\epsilon_t = \rho\epsilon_{t-1} + \eta_t,$$

he has shown that no uniformly most powerful or type B_1 test exists. From his arguments it appears that, of the statistics whose exact distribution is known, the statistics c_1 and d_3 should be most suitable respectively for series with a fixed mean, known and unknown.

As has been noted elsewhere (Durbin and Watson [3]), Anderson's results give us very little guidance for testing in a general regression model. Consequently the statistics suggested are justified only by their intuitive reasonableness. On the same intuitive grounds it is evident that the device of fitting two separate regressions is likely to bring about a substantial loss of power if the number of independent variables is not small compared with the number of observations.

The foregoing discussion of power has been conducted in terms of stationary Markov alternatives, partly because this is the case that is usually discussed, and partly because of its relative simplicity, not because we consider it to be of outstanding practical importance. For many cases found in practice the hypothesis that the errors follow a stationary stochastic process seems to us unrealistic. More usually, serial correlation of the errors will be due to systematic behaviour arising from the inadequacy of the theoretical model to represent the true relationship. This is a commonplace in econometrics, where tests of serial correlation are now often used as a routine procedure in the construction of models. In such situations the inappropriateness of a statistic which treats the products x_ix_{i-1} and x_1x_n on the same footing is evident on intuitive grounds. On the same grounds the statistics proposed above might prove to be more acceptable in many such cases.

5. Significance tables of the statistics c_1, d_3. In Section 4 the statistic c_1 was suggested for a test of randomness in a sample of an even number of observations from a population of known mean. A small table of the significance points of this statistic is given below. If the observed value of c_1 is greater than the tabulated value, the null hypothesis of randomness will be rejected at the 5% level of significance in favour of the hypothesis that positive serial correlation is present. As the distribution of c_1 is symmetrical about zero, a test for negative serial correlation may be made by considering $-c_1$. If the sample size is odd, the central observation could be dropped and the tests made as above with c_1.

5% point of c_1 for various n

10	12	14	16	18	20	22
0.426	0.403	0.382	0.364	0.348	0.334	0.321

For a test of serial independence in a series of unknown mean, the statistic d_3 has been suggested when the sample size is even. If the observed value of d_3 is less than the value tabulated below, the null hypothesis is rejected at the 5% level in a test against the alternative of positive serial correlation. For samples of odd size, the middle observations may be omitted so that d_3 is still applicable.

5% point of d_3 for various n

12	14	16	18	20	22	24	26	28	30
0.967	1.04	1.11	1.16	1.20	1.24	1.27	1.30	1.33	1.35

REFERENCES

[1] R. L. ANDERSON, "Distribution of the serial correlation coefficient," *Annals of Math. Stat.*, Vol. 13 (1942), pp. 1–13.

[2] T. W. ANDERSON, "On the theory of testing serial correlation," *Skandinavisk Aktuarietidskrift*, Vol. 31 (1948), pp. 88–116.

[3] J. DURBIN AND G. S. WATSON, "Testing for serial correlation in least squares regression. I," *Biometrika*, Vol. 37 (1950), pp. 409–428.

[4] B. I. HART, "Significance levels for the ratio of the mean square successive difference to the variance," *Annals of Math. Stat.*, Vol. 13 (1942), pp. 445–447.

[5] J. VON NEUMANN, "Distribution of the ratio of the mean square successive difference to the variance," *Annals of Math. Stat.*, Vol. 12 (1941), pp. 367–395.