On testing for serial correlation in large numbers of small samples

BY D. R. COX
Department of Mathematics, Imperial College, London SW7 2BZ, U.K.

AND P. J. SOLOMON
Department of Statistics, University of Adelaide, Adelaide, South Australia 5001, Australia

SUMMARY
Methods are outlined for testing for serial correlation in large numbers of small samples, emphasis being placed on samples of size three. Some data on pulse rates are analysed in illustration.

Some key words: Analysis of variability; Asymptotic relative efficiency; Normal theory; Panel data; Pulse rate; Quadratic form; Serial correlation.

1. INTRODUCTION
Recently Cox & Solomon (1986) examined a number of methods, formal and informal, for analysing sets of small samples in order to detect departures from the standard assumptions of normality and constant variance. The types of departure studied included common nonnormality and randomly changing variance. Serial correlation within samples was, however, mentioned only incidentally. In the present note, we consider testing for such correlation. The associated estimation problems are more delicate, particularly in view of the inconsistency of the simple maximum likelihood estimate, and will not be studied.

Suppose then that we have m independent samples each of r observations represented by random variables Y_{is} (i = 1,..., m; s = 1,..., r). As in the previous paper we suppose that E(Y_{is}) = \mu_i and that for the current analysis \mu_1,..., \mu_m are unknown nuisance parameters. Under the standard normal theory assumptions the \{Y_{is}\} are independently normally distributed with unknown constant variance \sigma^2. Here we concentrate on detecting serial correlation within the samples: as noted above, it is assumed that there is no systematic dependence of the mean on serial order. Where necessary to formulate a specific alternative involving serial correlation we shall suppose that, for each i, \{Y_{i1},..., Y_{ir}\} forms a stationary first-order autoregressive process of correlation \rho. We leave open the question of whether the different samples have the same variance.

For the most part we concentrate on r = 3.

2. DEVELOPMENT OF TEST STATISTIC
Consider first a single set of three random variables Y_1, Y_2, Y_3 of mean \bar{Y}. Although not the only possibility it is natural to base the analysis on the lag one sum of products

\[(Y_1 - \bar{Y})(Y_2 - \bar{Y}) + (Y_2 - \bar{Y})(Y_3 - \bar{Y}) = -(Y_2 - \bar{Y})^2,\]

and thus to calculate from a single sample

\[-(Y_2 - \bar{Y})^2/S,\]
where $S = \Sigma (Y_i - \bar{Y})^2$. That this is essentially negative is a consequence of the strong negative correlation between residuals in a sample of three.

To find the distribution of (1) when $\rho = 0$, that is when the observations are independent, we transform the column vector $Y$ by a permutation of the standard Helmert matrix, writing

$$
\begin{bmatrix}
Z_1 \\
Z_2 \\
Z_3
\end{bmatrix} = \begin{bmatrix}
1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\
-1/\sqrt{6} & 2/\sqrt{6} & 1/\sqrt{6} \\
1/\sqrt{2} & 0 & -1/\sqrt{2}
\end{bmatrix} \begin{bmatrix}
Y_1 \\
Y_2 \\
Y_3
\end{bmatrix},
$$

so that (1) becomes $-\frac{3}{2}Z_2^2/(Z_2^2 + Z_3^2)$. This suggests defining

$$V = 3(Y_2 - \bar{Y})^2/S = 2Z_2^2/(Z_2^2 + Z_3^2).$$

(2)

Under the null hypothesis $V$ has a beta-type density and

$$E(V) = 1, \quad \text{var} (V) = \frac{1}{2}.$$  

(3)

For possible use in graphical analysis, note that $\pi^{-1} \cos^{-1} (V - 1)$ is uniformly distributed on $(0, 1)$.

We now have at least two possible analyses for $m$ sets of data. One is to calculate $V_i$ for the $i$th set and:

(i) to use $\hat{V} = (\Sigma V_i/m)$ as a test statistic, approximately normally distributed with unit mean and variance $\frac{1}{2}m^{-1}$;

(ii) to make a cumulative plot of the ordered $V_i$ checking for uniformity of distribution.

The analysis in effect allows different sets of data to have different variances, $\sigma_i^2$.

A second possibility is to assume $\sigma_i^2$ to be constant. Then it is reasonable to replace $\Sigma V_i$ by

$$\hat{T} = 3 \sum (Y_{i2} - \bar{Y}_i)^2/\sum S_i = 2 \sum Z_{i2}^2/(Z_{i2}^2 + Z_{i3}^2),$$

(4)

i.e. to pool numerators and denominators. An exact test is obtained from (4) by noting that $F = \hat{T}/(2 - \hat{T})$ has the standard variance ratio distribution with $(m, m)$ degrees of freedom; asymptotically $\hat{T}$ is normal with mean one and variance $1/m$.

When $\sigma^2$ is constant, it is reasonable to expect $\hat{T}$ to be more efficient than $\hat{V}$. To study this quantitatively we need the local nonnull behaviour of $E(\hat{T})$ and $E(\hat{V})$, that is $\partial E(\hat{T}; \rho)/\partial \rho$ and $\partial E(\hat{V}; \rho)/\partial \rho$ at $\rho = 0$, in a self-explanatory notation.

3. NONNULL BEHAVIOUR

If $\{Y_{im}\}$ forms a stationary time series with autocorrelation $\rho_h$ of lag $h$, then

$$E\{(Y_{i2} - \bar{Y}_i)^2\} = \frac{3}{2}\sigma^2(1 - \frac{3}{4}\rho_1 - \frac{1}{3}\rho_2), \quad E(S_i) = 2\sigma^2(1 - \frac{3}{2}\rho_1 - \frac{1}{3}\rho_2),$$

so that $\hat{T}$ tends in probability to

$$1 - \frac{3}{2}\rho + O(\rho^2).$$

Now in the first-order autoregressive case $\rho_2 = \rho_1^2 = \rho^2$, say, and for small $\rho$ the statistic $\hat{T}$ therefore tends to

$$1 - \frac{3}{2}\rho + O(\rho^2).$$

(5)

It follows that the Pitman efficacy of $\hat{T}$, the square of the derivative of the expectation divided by the null variance, is $4m/9$.

The corresponding calculation for $\hat{V}$ is slightly more complicated. For a single sample from the first-order autoregressive process with $\mu = 0$, $\sigma = 1$, we can write

$$Y_1 = W_1, \quad Y_2 = \rho W_1 + W_2 + O_\rho(\rho^2), \quad Y_3 = \rho W_2 + W_3 + O_\rho(\rho^2),$$

where $W_1$, $W_2$, $W_3$ are independent standard normal random variables. We can now show that locally near $\rho = 0$

$$E(\hat{V}) = 1 - \frac{1}{2}\rho + O(\rho^2).$$

(6)
Thus the efficacy of $\tilde{V}$ is $2m/9$, showing that when the $\sigma_i^2$ are all equal the asymptotic relative efficiency of $\tilde{V}$ relative to $\tilde{T}$ is $\frac{1}{3}$.

These local results have been tested by a short simulation with $m = 10$, $p = 0$, $\pm \frac{1}{2}$, the reasonably close agreement between analytical and simulated results providing a quite severe test of the usefulness of these local approximations. To give a rough idea of sensitivity, note that to the approximations indicated, a two-sided 0.05 level test would have power $\frac{1}{2}$ at $\rho = \pm \frac{1}{2}$ for $m = 35$ if $\tilde{T}$ is used, and for $m = 138$ if $\tilde{V}$ is used.

4. Samples of Size Exceeding Three

We now deal more briefly with the corresponding analysis for $r > 3$. For a single set of random variables $Y_1, \ldots, Y_r$ of mean $\bar{Y}$, let

$$ Q = (Y_1 - \bar{Y})(Y_2 - \bar{Y}) + \ldots + (Y_{r-1} - \bar{Y})(Y_r - \bar{Y}) = \sum Y_i Y_{i+1} - (r+1)\bar{Y}^2 + \bar{Y}(Y_1 + Y_r), $$

$$ S = \sum (Y_i - \bar{Y})^2. $$

If for the $i$th set we denote the values of these statistics by $(Q_i, S_i)$, we concentrate here on $T' = \sum Q_i / \sum S_i$, although the relative efficiency of a procedure using instead $\sum (Q_i / S_i)$ will presumably increase with $r$.

To obtain the asymptotic distribution of the test statistic under the null hypothesis, it is enough to find the first and second moments of $(Q, S)$ when $Y_1, \ldots, Y_r$ are independent and identically distributed in the standard normal distribution. These moments are most systematically calculated via results for quadratic forms in standard normal variables (Searle, 1971, p. 57), namely

$$ E(Y^TAY) = \text{tr}(A), \quad \text{var}(Y^TAY) = 2 \text{tr}(A^2), \quad \text{cov}(Y^TAY, Y^TBY) = 2 \text{tr}(AB), $$

where $A, B$ are arbitrary constant matrices. On identifying the matrices associated with $Q$ and $S$, we find after appreciable calculation that

$$ E(Q) = -(r-1)/r, \quad \text{var}(Q) = (r^2-3r^2+2r+2)/r^2, $$

$$ E(S) = r-1, \quad \text{var}(S) = 2(r-1), \quad \text{cov}(Q, S) = -2(r-1)/r. $$

It follows after some further calculation that under the null hypothesis the approximate mean and variance of the limiting normal distribution of $T'$ are

$$ -1/r, \quad \{(r+1)(r-2)^2\}/\{mr^2(r-1)^2\}, $$

agreeing with the results given previously for $r = 3$.

5. Example

As a brief illustrative example, we use, as in our previous paper, triple observations on pulse rate for samples of 100 men and 100 women from IPPPSH (International Primary Prevention Study on Hypertension), giving only outline conclusions because of constraints on space. The nature of the data is explained in the earlier paper. Unfortunately, for geographical reasons, associated with the authors not with the patients, the set of patients used here is not identical with that used previously, although there is substantial overlap.

The previous analysis had suggested the presence of overdispersion, somewhat reduced but not eliminated by taking reciprocals, and inspection of the data indicated some digit preferences and rounding. The analysis using $\tilde{V}$ thus seemed preferable to that using $\tilde{T}$, which in effect assumes constant variance, although both analyses were performed. All analyses were done separately for men and women and were done with and without the reciprocal transformation.

The distributions of $\pi^{-1} \cos^{-1}(V_i - 1)$ were nonuniform but rather irregular; overinterpretation of details of the plots is hazardous in the presence of the complications like rounding mentioned above.
Table 1 summarizes the statistics $\hat{V}$ and $\hat{T}$. Agreement between the values for males and females is remarkably close. As would be expected, transformation has little effect on the statistics, since qualitatively serial correlation is unaffected by such transformation. Had the statistic become, say, more extreme after transformation, that would have suggested more nearly linear correlation on the transformed scale. The statistics all point towards modest positive serial correlation, the more cautious $\hat{V}$ test showing that, while separately for males and females the statistics are barely significant, nevertheless the combined evidence is quite compelling.

Table 1. Sets of three pulse rates for 100 men and 100 women, IPPPSH. Test statistics for serial correlation

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Untransformed Reciprocal</td>
<td>Untransformed Reciprocal</td>
</tr>
<tr>
<td>$\hat{V}$</td>
<td>0.861 0.877</td>
<td>0.862 0.901</td>
</tr>
<tr>
<td>$\hat{T}$</td>
<td>0.898 0.847</td>
<td>0.811 0.831</td>
</tr>
</tbody>
</table>

Under null hypothesis $\hat{V}$ and $\hat{T}$ have unit mean and standard errors 0.0707 and 0.1.

Acknowledgements

We thank Ciba-Geigy Ltd, Basel for permission to use data from International Prospective Primary Prevention Study on Hypertension. D. R. Cox is grateful to Science and Engineering Research Council for a Senior Research Fellowship.

References


[Received February 1987. Revised May 1987]