551·513·7 : 551·509·33

Estimating the bias of correlation coefficients

ALISON M. GRANT

Department of Meteorology, University of Melbourne, Australia

(Received 11 November 1952)

1. Introduction

One of the major problems arising in the correlation method of seasonal forecasting is the selection of factors to be used in forecast formulae. As pointed out by Savur (1935) there are two main ways in which the choice can be made—

(*i*) a priori selection in which the factors are chosen on theoretical or physical grounds, and

(*ii*) selection by correlation in which the factors are chosen because empirical relationships have been found to exist between them and the elements for which forecasts are required.

It was early recognized that in testing the significance of correlations selected under (*ii*) some allowance should be made for the fact that they have in general been chosen as the highest of a large number. For this purpose Walker (1914) gave values of the "probable highest" correlation while Savur and Gopal Rao (1932) extended this to give values necessary for significance at the 5 per cent level. However, effects of selection should also be taken into account when estimating the true or long term value of a correlation coefficient, a matter which is of even greater concern since the best estimate is required to determine the actual regression equation.

In examining effects of selection from this point of view it is simplest to deal with the transformed correlation z* which is very nearly normally distributed about the value ζ (similarly defined from the population correlation ρ) with variance $1/(n-3)$, where n is the number of pairs of values on which the sample correlation r is based. Corresponding to each calculated correlation we may write $z = \zeta + d$ where d represents the sampling error. A possibility which cannot be ignored is that in selecting the highest values of z one is in effect choosing those which have high values of d. If this were the case some idea of the bias associated with the i^{th} highest value of z could be gained from the mean of the i^{th} highest value of d. The present note is concerned with a simple method of obtaining the mean of the i^{th} highest value in a sample of size m from a standard normal distribution, and with the application of this result to selected correlation coefficients.

2. Mean of the i^{th} highest value

The probability that the i^{th} highest out of a sample of m values from a standard normal distribution is between x and $x + dx$ is equal to the probability of obtaining $m-i$ values below x, i-1 values above $x + dx$, and one value between x and $x + dx$. In the limit this is given by

$$\frac{m\,!}{(m-i)\,!\,(i-1)\,!}\,p^{m-i}\,q^{i-1}\,\frac{1}{(2\pi)^{\frac{1}{2}}}\,e^{-\frac{1}{2}x^2}\,dx$$

where p is the probability of any particular value being below x, *i.e.*,

$$p = \frac{1}{(2\pi)^{\frac{1}{2}}}\int_{-\infty}^{x} e^{-\frac{1}{2}x^2}\,dx$$

and $q = 1-p$. The mean of this distribution can be evaluated for given values of i and m only by a tedious numerical process such as was carried out by Tippett (1925) for $i = 1$ and certain values of m

* $z = \frac{1}{2}\left\{\log(1+r) - \log(1-r)\right\}$

An approximate value for the mean can be found by using the median, which is obtained by solving the equation

$$p^m + \binom{m}{1} p^{m-1} q + \binom{m}{2} p^{m-2} q^2 + \ldots\ldots + \binom{m}{i-1} p^{m-i+1} q^{i-1} = \tfrac{1}{2}$$

for p and finding the corresponding value of x from standard normal tables. This method was used by Savur (1935) to give values of correlations corresponding to Walker's 50 per cent level of significance. The same values (denoted by ρ_w in his paper) were then used as estimates of the true correlation which strictly speaking involved the use of the median in place of the mean. This method has the disadvantage that for $i \neq 1$ the equation giving the median value of p requires solution by numerical methods.

However it will now be shown that the mean of the i^{th} highest value in a sample of size m from a standard normal distribution can be approximated by another quantity which is much easier to compute than the median used by Savur (1935). This quantity, denoted by $X_{i,m}$, is defined so that on the average $i - \frac{1}{2}$ of the m values in the sample would exceed it, *i.e.*, so that

$$\frac{1}{(2\pi)^{\frac{1}{2}}} \int_{X_{i,m}}^{\infty} e^{-\frac{1}{2}x^2} dx = \frac{i - \frac{1}{2}}{m}$$

A comparison of values of $X_{i,m}$ with the corresponding values of the mean (from Tippett 1925) and the median are given in Table 1, for $i=1$ and certain values of m.

TABLE 1

Mean of the highest value in a sample of size m from a standard normal distribution

m	Exact Value Tippett (1925)	Approximate values Median Savur (1935)	 $X_{i,m}$
2	0·56	0·54	0·67
5	1·16	1·13	1·28
10	1·54	1·50	1·64
20	1·87	1·82	1·96
60	2·32	2·27	2·39
100	2·51	2·46	2·58
200	2·75	2·70	2·81
500	3·04	2·99	3·09
1000	3·24	3·20	3·29

The figures show that $X_{i,m}$ is almost as good an approximation to the mean as is the median. Moreover the error made in either case is very much smaller than errors which must be expected due to sampling since the standard deviation of the highest value (also given by Tippett 1925) varies from 0·83 for $m = 2$ to 0·35 for $m = 1000$. In no case does the error due to the approximation exceed 1/5 of the corresponding standard deviation so that for practical purposes either the median or the value $X_{i,m}$ may be used in place of the mean.

For $i \neq 1$ a direct comparison cannot be made as the exact mean values are not available. However the differences would probably be smaller than those shown for $i = 1$ since the distributions become more symmetrical as more central values are considered. For example for the central value when m is odd $[i = \frac{1}{2}(m+1)]$ mean, median, and $X_{i,m}$ are all zero. It, therefore, seems likely that the value $X_{i,m}$ would be a good approximation to the mean for $i \neq 1$, a result which is of particular interest since the median is much more difficult to obtain for higher values of i.

In order to simplify the practical application of the suggested approximation, values of $X_{i,m}$ for different values of i and m are given in Table 2.

3. Application to Selected Correlation Coefficients

If we consider m correlations each obtained from n pairs of values, the bias associated with the transformed correlation z which corresponds to the i^{th} highest value of the sampling error d, is given by

$$\frac{X_{i,m}}{(n-3)^{\frac{1}{2}}}$$

TABLE 2

Approximate mean of the i^{th} highest value in a sample of size m from a standard normal distribution ; Values of $X_{i,m}$

i m	1	2	3	4	5	6	7	8	9	10
1	0									
2	0·7									
3	1·0	0								
4	1·2	0·3								
5	1·3	0·5	0							
6	1·4	0·7	0·2							
7	1·5	0·8	0·4	0						
8	1·5	0·9	0·5	0·2						
9	1·6	1·0	0·6	0·3	0					
10	1·6	1·0	0·7	0·4	0·1	—0·1				
12	1·7	1·2	0·8	0·5	0·3	0·1	—0·1			
14	1·8	1·2	0·9	0·7	0·5	0·3	0·1	—0·1		
16	1·9	1·3	1·0	0·8	0·6	0·4	0·2	0·1	—0·1	
18	1·9	1·4	1·1	0·9	0·7	0·5	0·4	0·2	0·1	—0·1
20	2·0	1·4	1·2	0·9	0·8	0·6	0·5	0·3	0·2	0·1
30	2·1	1·6	1·4	1·2	1·0	0·9	0·8	0·7	0·6	0·5
50	2·3	1·9	1·6	1·5	1·3	1·2	1·1	1·0	1·0	0·9
100	2·6	2·2	2·0	1·8	1·7	1·6	1·5	1·4	1·4	1·3
500	3·1	2·7	2·6	2·5	2·4	2·3	2·2	2·2	2·1	2·1
1000	3·3	3·0	2·8	2·7	2·6	2·5	2·5	2·4	2·4	2·3

i m	10	12	14	16	18	20	30	50	100
20	0·1	—0·2	—0·5						
30	0·5	0·3	0·1	—0·1	—0·3	—0·5			
50	0·9	0·7	0·6	0·5	0·4	0·3	—0·2	—2·3	
100	1·3	1·2	1·1	1·0	0·9	0·9	0·5	0	—2·6
500	2·1	2·0	1·9	1·9	1·8	1·8	1·6	1·3	0·8
1000	2·3	2·3	2·2	2·2	2·1	2·1	1·9	1·6	1·3

where $X_{i,m}$ is obtained from Table 2 with the given values of i and m. It is of course impossible to identify any particular observed correlation with a particular value of i. If, however, the k highest correlations have between them most of the high sampling errors, the *average* bias will not be very different from that derived on the assumption that the i^{th} highest error is associated with the i^{th} highest observed correlation. For example the 5 highest correlations out of 50 might in reality have sampling errors such that $i = 4, 1, 6, 2, 10$. The average value of $X_{i,m}$ for $m = 50$ and these values of i is found from Table 2 to be 1·6, while that for $i = 1, 2, 3, 4, 5$ is equal to 1·7, It is, therefore, suggested that each of the k transformed correlations be reduced by an amount

$$\frac{Y_{k,m}}{(n-3)^{\frac{1}{2}}}$$

where $Y_{k,m}$ is the mean value of $X_{i,m}$ for

$i = 1, 2, \ldots, k$; *i.e.*,

$$Y_{k,m} = \frac{1}{k} \sum_{i=1}^{k} X_{i,m}$$

This will lead on the average to unbiassed estimates of the transformed correlations which may be used to determine corrected correlation coefficients.

The relevance of this procedure depends on the fulfilment of the condition mentioned above, *viz.*, that the k highest correlations have between them most of the high errors. Ultimately this must be decided by reference to selected correlations themselves, in particular by comparing the theoretical mean bias with the average drop observed in the transformed correlations when independent estimates are available. A study was, therefore, made of the relationships used in actual forecast formulae. As an example the results obtained with reference to the formulae for Indian rainfall published by Walker (1924) are set out below.

From the information given it was estimated that about 500 relationships between past and future weather were considered in the derivation of the six forecast formulae. Of the 28 relationships finally used, Savur (1935) regarded 8 as being selected on theoretical grounds, so that the remaining 20 were in effect chosen as the highest of about 500. Averaging the values of $X_{i,m}$ (from Table 2) for $i = 1, 2, \ldots, 20$ and $m = 500$ gives $Y_{k,m} = 2 \cdot 1$, while the average period on which the original correlations were based is 31 years. The mean theoretical bias is, therefore, given by

$$\frac{Y_{k,m}}{(n-3)^{\frac{1}{2}}} = \frac{2 \cdot 1}{(28)^{\frac{1}{2}}} = 0 \cdot 40$$

Independent estimates of these same 20 correlations were given by Savur (1935) in Table 1 of his paper. From these values the average drop observed in the transformed correlations was computed and found to be 0·39, showing that the selected correlations do in fact behave as though they were associated with the highest 20 out of 500 errors.

The assumption underlying the theoretical mean bias is, therefore, justified with regard to Walker's (1924) correlations and it seems not unlikely that the suggested procedure should be used generally for correlation coefficients selected because of their magnitude.

REFERENCES

Fisher, R. A. (1932). *Statistical Methods for Research Workers*, 4th edition (Oliver and Boyd), p. 176.

Savur, S. R. (1935). *Sankhya*, **2**, Pt. 1, pp. 2-12.

Savur, S. R. and Gopal Rao, S. (1932). *Ind. met. Dep. Sci. Notes*, **5**, 49.

Tippett, L. H. C. (1925). *Biometrika*, **17**, pp. 364-387.

Walker, G. T. (1914). *Mem. Ind. met. Dep.*, **21**, Pt. 9, pp. 13-15.

Walker, G. T. (1924). *Mem. Ind. met. Dep.*, **24**, Pt. 10, pp. 333-345.