ARKIV FÖR MATEMATIK Band 3 nr 31

Communicated 23 May 1956 by HARALD CRAMÉR

On linear estimates with nearly minimum variance

By GUNNAR BLOM

1. Introduction

Let z be a random variable with a continuous cumulative distribution function $F[(z-\mu)/\sigma]$ which depends upon two unknown parameters μ and σ. Consider an ordered random sample

$$z_{(1)} \leqq z_{(2)} \leqq \cdots \leqq z_{(n)}$$

of z-values. If the means and covariances of the reduced order statistics

$$x_i = \frac{z_{(i)} - \mu}{\sigma} \qquad (i = 1, 2, \ldots, n)$$

are known, it is possible to find linear unbiased minimum variance estimates

$$\sum_{i=1}^{n} g_{1i} z_{(i)} \quad \text{and} \quad \sum_{i=1}^{n} g_{2i} z_{(i)}$$

of μ and σ respectively (Lloyd, 1952). These estimates may be called best unbiased estimates. A serious drawback of the solution is that in most cases it involves very time-consuming numerical calculations.

The object of this paper is to show that, under general conditions, it is possible to find a convenient approximation to the best solution which may be termed a nearly best unbiased estimate. The variance of this estimate is, as some examples will show, often very little in excess of the minimum variance. The method presupposes that the means (but not the covariances) of the variables x_i are known.

By a slight modification of the method it may be used also when neither the means nor the covariances are known. The resulting estimates will be called nearly best, nearly unbiased estimates.

Both types of estimates mentioned above may be derived from a theorem given in the next section.

2. A theorem on linear estimates

Denote by $E x_i$ and $cov\,(x_i, x_j)$ the means and covariances of the variables x_i $(i = 1, 2, \ldots, n)$. Put $p_i = i/(n+1)$ and $q_i = 1 - p_i$. Further, define $\lambda_i = G(p_i)$, where $G(u)$ is the inverse function of $F(x)$.

Consider a linear estimate

$$U = \sum_{i=1}^{n} g_i z_{(i)} \tag{2.1}$$

of the parameter $k_1 \mu + k_2 \sigma$, where k_1 and k_2 are given constants.

Theorem. *Let*

$$\theta_0, \theta_1, \ldots, \theta_n, \theta_{n+1}$$

be any set of quantities subject to the conditions

$$\theta_0 = \theta_{n+1} = 0 \quad \textit{and} \quad \theta_i \neq 0 \quad (i = 1, 2, \ldots, n).$$

Replace the coefficients g_i in (2.1) *by new quantities*

$$h_0, h_1, \ldots, h_n$$

defined, apart from an additive constant, by the relations

$$g_i = \theta_i (h_i - h_{i-1}) \qquad (i = 1, \ldots, n). \tag{2.2}$$

If the conditions

$$\sum_{i=0}^{n} C_{1i} h_i = k_1 ; \qquad \sum_{i=0}^{n} C_{2i} h_i = k_2 \tag{2.3}$$

are fulfilled, where

$$C_{1i} = \theta_i - \theta_{i+1} ; \qquad C_{2i} = \theta_i E x_i - \theta_{i+1} E x_{i+1} \qquad (i = 0, 1, \ldots, n) \tag{2.4}$$

then U in (2.1) *is an unbiased estimate of $k_1 \mu + k_2 \sigma$.*

Further, the variance of U can be written

$$\textit{var}\ U = \frac{\sigma^2}{n+2} \left[\frac{1}{n+1} \sum_{i=0}^{n} h_i^2 - \left(\frac{1}{n+1} \sum_{i=0}^{n} h_i \right)^2 \right] + \sigma^2 R, \tag{2.5}$$

where

$$R = \sum_{i,j=1}^{n} g_i g_j R_{ij},$$

R_{ij} being defined by

$$\textit{cov}\ (x_i, x_j) = \frac{p_i q_j}{n+2} (\theta_i \theta_j)^{-1} + R_{ij} \qquad (i \leqq j).$$

The theorem holds for any choice of quantities θ_i. It is, however, useful only when they are chosen so that R in (2.5) behaves in a suitable way. In particular, it is desirable to choose θ_i so that R tends rapidly to zero when n increases. When $F(x)$ has a derivative $f(x)$ which is continuous (except, possibly, in the end-points of the range of variation), we take $\theta_i = f(\lambda_i)$.

3. Derivation of the nearly best unbiased estimate

The variance of U given in (2.5) is, apart from the last term, proportional to

$$Z = \sum_{i=0}^{n} h_i^2 - \frac{1}{n+1}\left(\sum_{i=0}^{n} h_i\right)^2.$$

Let now Z be minimized with respect to the h_i subject to the side conditions (2.3). The resulting quantities h_i obviously provide an unbiased estimate of $k_1\mu + k_2\sigma$, but there is no guarantee that the true minimum of *var* U is obtained, since R is neglected. Determining h_i in this way and using (2.2) we find as coefficients of the nearly best estimate

$$g_i = \theta_i[a_1(C_{1i} - C_{1i-1}) + a_2(C_{2i} - C_{2i-1})], \tag{3.1}$$

where a_1 and a_2 are two multipliers. Introducing the notation

$$d_{\mu\nu} = \sum_{i=0}^{n} C_{\mu i} C_{\nu i} \qquad (\mu, \nu = 1, 2),$$

$$D = d_{11} d_{22} - d_{12}^2,$$

we have

$$a_1 = \frac{1}{D}(k_1 d_{22} - k_2 d_{12}),$$

$$a_2 = \frac{1}{D}(-k_1 d_{21} + k_2 d_{11}).$$

When the distribution is symmetrical, $d_{12} = d_{21} = 0$.

The coefficients of the nearly best estimates of μ and σ are obtained by taking $k_1 = 1$, $k_2 = 0$ and $k_1 = 0$, $k_2 = 1$ respectively in these relations. It should be noticed that the solution depends upon the first and second order differences of the sequences $\{\theta_i\}$ and $\{\theta_i E x_i\}$ $(i = 0, 1, \ldots, n+1)$.

The nearly best estimates of μ and σ have been determined in the case $n = 5$ for some special functions $F(x)$ for which best estimates are known. In these examples the quantities θ_i have been put equal to $f(\lambda_i)$.

As a measure of the efficiency of the estimates the quotient of the variance of the best estimate and the variance of the nearly best estimate has been used. (The last-mentioned quantity has been determined accurately by aid of existing tables, *not* from the approximation formula obtained by neglecting R in (2.5).) The result of the calculations is given in Table 1. As seen from the table the loss of efficiency by using nearly best instead of best estimates is very small in these examples. (The estimate of μ in the normal case is of little interest but has been included for completeness.)

It should be mentioned that, in the rectangular case, the nearly best estimate is identical with the well-known best estimate based upon the extreme values of the sample.

When $F(x)$ has a continuous derivative which vanishes in the end-points of the range of variation, it may sometimes be convenient to replace the second

Table 1. Efficiency of nearly best linear estimates.

Distribution	Parameter		Best estimate given by
	μ	σ	
Rectangular	100 %	100 %	Lloyd
Normal	99.8	99.7	Godwin
Triangular	94.4	99.6	Sarhan
Extreme value	98.8	96.6	Lieblein
Right triangular	99.9	99.8	Downton

Efficiency = quotient of variance of best linear unbiased estimate and variance of nearly best linear unbiased estimate.

order differences in (3.1) by the derivatives of $f(x)$. If this modification is used in the case of the normal distribution, the nearly best estimate of μ is replaced by the sample mean and the nearly best estimate of σ by the estimate

$$\frac{\sum_{i=1}^{n} \lambda_i z_{(i)}}{\sum_{i=1}^{n} \lambda_i E x_i}.$$

The variance of this estimate is practically the same as that of the best estimate, the loss of efficiency being about 0.1 % when $n = 5$.

It deserves to be mentioned that the estimates obtained by the modified method are, apart from the values of the multipliers, identical with the estimates studied by Jung (1955) in the case of a class of distributions which, however, does not include the normal distribution.

4. Nearly best, nearly unbiased estimates

When the means of the variables x_i are unknown, the approximation

$$E x_i \sim G\left(\frac{i-\alpha}{n-\alpha-\beta+1}\right)$$

is used in the definition of C_{2i} in (2.4). In other respects the procedure is the same as in the preceding section. The resulting estimates will in general be biased. By a suitable choice of constants α and β the bias may, however, often be very small. General rules for the selection of the constants may be formulated. The variance of the estimates obtained in this way is often practically the same as that of the nearly best unbiased estimate.

Finally it deserves to be mentioned that by a study of the asymptotic variance of nearly best estimates, it is possible to obtain information about the nature of estimates which are non-regular in the sense used by Cramér (1946, Ch. 32).

Details concerning the method described in this paper will be given in a forthcoming publication.

REFERENCES

CRAMÉR, H. (1946), Mathematical Methods of Statistics. Princeton.
DOWNTON, F. (1954), Least-squares estimates using ordered observations. Ann. Math. Stat. *25*, 303.
GODWIN, H. J. (1949), On the estimation of dispersion by linear systematic statistics. Biometrika *36*, 92.
JUNG, J. (1955), On linear estimates defined by a continuous weight function. Arkiv för Matematik Bd *3* nr *15*, 199.
LIEBLEIN, J. (1954), A new method of analyzing extreme-value data. Nat. Adv. Comm. Aeronat. Techn. Note 3053.
LLOYD, E. H. (1952), Least-squares estimation of location and scale parameters using order statistics. Biometrika *39*, 88.
SARHAN, A. E. (1954), Estimation of the mean and standard deviation by order statistics. Ann. Math. Stat. *25*, 317.

Tryckt den 27 december 1956

Uppsala 1956. Almqvist & Wiksells Boktryckeri AB