The Annals of Mathematical Statistics
1971, Vol. 42, No. 3, 1084-1087

SHORT COMMUNICATIONS

ON SPLITTING A SYSTEMATIC SAMPLE FOR VARIANCE ESTIMATION

BY J. C. KOOP

Dominion Bureau of Statistics, Ottawa

0. Summary. Variance estimation in systematic sampling by splitting the sample into equal halves can lead to very serious bias. The expression for this bias relative to the true variance is given in terms of intraclass correlation coefficients. The danger of serious bias is still present when successive pairs of units are treated as "independent" replicates; an expression for this relative bias is also given.

1. Introduction. Sometimes a systematic sample (say s) of n units, drawn from a universe of $N = nk$ units, where n is even, is split up into two systematic subsamples A and B for the purpose of estimating the variance of the mean of the sample, $\bar{x}$. To exclude trivial considerations we state that at least two units of the universe have variate values which are unequal so that the variance $\sigma^2 > 0$, and further the sample size $n > 2$.

Let $x_{i1}, x_{i2}, \cdots, x_{in}$ $(i = 1, 2, \cdots, k)$ be the variate values of the ith selected systematic sample. Subsample A is made up of all the even members of the sample and B of all the odd members. If $\bar{x}_A$ and $\bar{x}_B$ are the means of these subsamples then obviously $\bar{x} \equiv \bar{x}_{i.} = \sum_{j=1}^{n} x_{ij}/n = \frac{1}{2}(\bar{x}_A + \bar{x}_B)$. A question which arises is whether or not, on analogy with independent replicated sampling, $\frac{1}{4}(\bar{x}_A - \bar{x}_B)^2$ is an unbiased estimate of the variance of $\bar{x}$.

Treating the $\frac{1}{2}n$ successive pairs of units of the systematic sample as "independent" replicates, and computing their respective means, namely

$$\bar{x}_{it} = \tfrac{1}{2}(x_{i(2t-1)} + x_{i(2t)}), \qquad (t = 1, 2, \cdots, \tfrac{1}{2}n); \tag{1}$$

it is also interesting to enquire, as will appear in sequel, whether or not

$$v(\bar{x}) = \textstyle\sum_{t=1}^{\frac{1}{2}n} (\bar{x}_{it} - \bar{x}_{i.})^2 / \{\frac{1}{2}n(\frac{1}{2}n - 1)\} \tag{2}$$

is an unbiased estimate of the variance of $\bar{x}$.

We seek answers to these two questions in this paper.

2. Solution of problem when sample is split into halves. By definition

$$V(\bar{x}) = \tfrac{1}{4}\{V(\bar{x}_A) + V(\bar{x}_B) + 2\,\mathrm{Cov}\,(\bar{x}_A, \bar{x}_B)\}. \tag{3}$$

Noting that A and B are the only possible systematic subsamples from s, we find

$$V(\bar{x}_A) = V\{E(\bar{x}_A \mid s)\} + E\{V(\bar{x}_A \mid s)\}. \tag{4}$$

Received August 11, 1969; revised October 8, 1969.

Again by definition we find in (4)

$$E(\bar{x}_A \mid s) = \tfrac{1}{2}(\bar{x}_A+\bar{x}_B) = \bar{x}, \tag{5}$$

and

$$V(\bar{x}_A \mid s) = \tfrac{1}{2}\{(\bar{x}_A-\bar{x})^2+(\bar{x}_B-\bar{x})^2\} = \tfrac{1}{4}(\bar{x}_A-\bar{x}_B)^2. \tag{6}$$

With these results

$$V(\bar{x}_A) = V(\bar{x})+\tfrac{1}{4}E(\bar{x}_A-\bar{x}_B)^2. \tag{7}$$

Similarly it can be shown that $V(\bar{x}_B)$ is equal to the expression on the right-hand side of (7).

Defining ρ_0 to be the correlation between the means of systematic subsamples, i.e.,

$$\operatorname{Cov}(\bar{x}_A, \bar{x}_B) = \rho_0 V(\bar{x}_A) = \rho_0 V(\bar{x}_B), \tag{8}$$

in the context of the particular arrangement of the nk units from which the systematic sample s was drawn, we find that (3) can be rewritten

$$V(\bar{x}) = \tfrac{1}{2}(1+\rho_0)V(\bar{x}_A). \tag{9}$$

Eliminating $V(\bar{x}_A)$ between (9) and (7), we find, after rearrangement of terms,

$$\tfrac{1}{4}E(\bar{x}_A-\bar{x}_B)^2 = \left(1-\frac{2\rho_0}{1+\rho_0}\right)V(\bar{x}). \tag{10}$$

Clearly therefore $\frac{1}{4}(\bar{x}_A-\bar{x}_B)^2$ is a biased estimate of $V(\bar{x})$. Theoretically, if this estimate is multiplied by $(1+\rho_0)/(1-\rho_0)$ the result will be unbiased. But the suggestion is useless as we do not know ρ_0.

The relative bias in estimating $V(\bar{x})$, as indicated by (10), is $-2\rho_0/(1+\rho_0)$. We shall attempt to find an expression for ρ_0, and this relative bias, in terms of the intraclass correlations specific to the set of k particular systematic samples and the set of $2k$ particular subsamples resulting from splitting each possible sample into two halves, one half being made up of the even members and the other half of the odd members.

Let ρ and ρ' be the intraclass correlations specific to the sets of k and $2k$ systematic samples. Then from theory which is well known

$$V(\bar{x}) = \frac{\sigma^2}{n}\{1+(n-1)\rho\}, \tag{11}$$

and

$$V(\bar{x}_A) = V(\bar{x}_B) = \frac{2\sigma^2}{n}\left\{1+\left(\frac{n}{2}-1\right)\rho'\right\}. \tag{12}$$

We note that in view of (11) and (12)

$$-1/(n-1) < \rho < 1 \quad\text{and}\quad -2/(n-2) < \rho' < 1. \tag{13}$$

An examination of the underlying formulas for ρ and ρ' will show that the relative signs and magnitudes of these parameters will depend on the particulaı arrangement of the nk units. However, an inspection of the two inequalities given by (13) shows that it is quite possible for ρ' to be less than ρ even when they are both negative. From (9), (11) and (12) we find

$$\text{(14)} \qquad \rho_0 = \{1+(n-1)\rho\}/\{1+(\tfrac{1}{2}n-1)\rho'\}-1$$

so that the relative bias in estimating variance is

$$\text{(15)} \qquad (-2\rho_0)/(1+\rho_0) = \{(n-2)\rho'-2(n-1)\rho\}/\{1+(n-1)\rho\}.$$

For example when $n = 100$ and $\rho = \rho' = \cdot 01$, (15) shows that the variance will be underestimated by nearly 50%. Trivially also the bias is zero when $\rho = \rho' = 0$.

In general (15) shows that the variance will be overestimated, estimated without bias, or underestimated according as

$$\text{(16)} \qquad (\rho/\rho') \lesseqgtr (n-2)/2(n-1),$$

which is a positive quantity close to a half. From (16) we see that the necessary condition for the bias in the estimation of variance to be zero is that both ρ and ρ' must be of the same sign.

More generally when the sample is split into more than two systematic subsamples the conclusions are essentially the same. We omit the generalizations of (10), (14) and (15) to save space.

3. Splitting of sample by pairing successive units. Let $\bar{X}$ be the mean of the $N = nk$ units of the universe. Then it is easy to see that

$$\text{(17)} \qquad \bar{x}_{i.}-\bar{X} = \textstyle\sum_{t=1}^{\frac{1}{2}n} (\bar{x}_{it}-\bar{x}_{i.}+\bar{x}_{i.}-\bar{X})/(\tfrac{1}{2}n).$$

Squaring both sides of this identity and rearranging terms, we find

$$\text{(18)} \qquad (\bar{x}_{i.}-\bar{X})^2 = v(\bar{x})+\frac{8}{n(n-2)}\textstyle\sum_{t>t'} (\bar{x}_{it}-\bar{X})(\bar{x}_{it'}-\bar{X})$$

where $v(\bar{x})$ is as given by (2). We note that there are $n(n-2)/8$ cross-product terms on the right-hand side of (18). Remembering that the probability of realizing each systematic sample is $1/k$, and taking the expected values of expressions in (18) we find by definition

$$\text{(19)} \qquad E(\bar{x}_{i.}-\bar{X})^2 = V(\bar{x}) = E\{v(\bar{x})\}+\frac{8}{kn(n-2)}\textstyle\sum_{i=1}^{k}\sum_{t>t'} (\bar{x}_{it}-\bar{X})(\bar{x}_{it'}-\bar{X}).$$

Clearly therefore $v(\bar{x})$ is a biased estimate of the variance of $\bar{x}$, the amount of bias being equal to the cross-product term in (19) which will be denoted by C. We shall seek a more meaningful expression for this bias relative to the true variance.

Explicitly, with the use of (1), we find

$$C = \frac{2}{kn(n-2)} \textstyle\sum_{i=1}^{k} \sum_{t>t'} \{(x_{i(2t-1)} - \bar{X})(x_{i(2t'-1)} - \bar{X}) + (x_{i(2t)} - \bar{X}) \times (x_{i(2t'-1)} - \bar{X}) + (x_{i(2t-1)} - \bar{X})(x_{i(2t')} - \bar{X}) + (x_{i(2t)} - \bar{X})(x_{i(2t')} - \bar{X})\}. \tag{20}$$

It is not difficult to verify that the number of cross-product terms under the summation sign in (20) is exactly $kn(n-2)/2$. Therefore C may be regarded as a measure of intrapair covariation among the set of k systematic samples as will be evident from the formula itself. If we divide C by σ^2, we shall obtain some type of correlation coefficient, say $\tilde{\rho}$, which can be interpreted in the sense of the previous statement. Then by definition,

$$\tilde{\rho}\sigma^2 = C. \tag{21}$$

By the use of the Cauchy inequality, it is not difficult to show that $-1 < \tilde{\rho} < 1$.

With (21) and the formula for $V(\bar{x})$ given by (9), (19) reduces to

$$E\{v(\bar{x})\} = V(\bar{x})[1-(n\tilde{\rho})/\{1+(n-1)\rho\}]. \tag{22}$$

Hence the relative bias in using $v(\bar{x})$ to estimate the variance of $\bar{x}$ is

$$-n\tilde{\rho}/\{1+(n-1)\rho\}, \tag{23}$$

a result surprisingly similar to that obtained by setting $\rho' = \rho$ in the analogous formula given by (15). For example, when $\rho = \tilde{\rho} = \cdot 01$ and $n = 100$, the variance is underestimated by nearly 50%. Trivially the bias is zero if $\tilde{\rho} = 0$. Generally when ρ and $\tilde{\rho}$ are both equal and positive, underestimation increases with increasing n. It seems that the pairing of successive units is as dangerous as splitting the sample systematically into an equal number of parts.

4. Concluding remarks. To end on a constructive note it may be remarked that the difficulty of estimating variance without bias may be overcome by drawing two systematic samples each of size $\frac{1}{2}n$, as suggested by Madow and Madow (1944) [1], instead of drawing one solid systematic sample of size n. In this situation if $\bar{x}_A'$ and $\bar{x}_B'$ are the means of the samples, then an unbiased estimate of the variance of $\bar{x} = \frac{1}{2}(\bar{x}_A' + \bar{x}_B')$ will be

$$\tfrac{1}{4}(\bar{x}_A' - \bar{x}_B')^2 \left(\frac{k-1}{k}\right). \tag{24}$$

Acknowledgment. The writer is grateful to the referee for a suggestion which resulted in the theory presented in Section 3.

REFERENCE

[1] Madow, W. G. and Madow, L. H. (1944). On the theory of systematic sampling, I. *Ann. Math. Statist.* **15** 1–25.