A NOTE ON THE RATIO AND REGRESSION METHODS OF ESTIMATION IN CONTROLLED SIMPLE RANDOM SAMPLING

BY M. S. AVADHANI AND B. V. SUKHATME

Institute of Agricultural Research Statistics, New Delhi

1. *Introduction and Summary.* The authors have introduced the method of controlled simple random sampling, (*c f* [1]) to eliminate or reduce the chance of selection of non-preferred samples from the population and have shown that this technique would in general provide *no less* efficient estimator for the population total than the usual technique of simple random sampling. But when a simple random sample is taken for a survey inquiry and if ancillary information is available on all the units of the population, then it is customary to utilise this knowledge, by adopting either the ratio, or regression, method of estimation, for providing more efficient estimate for the population total. Therefore it would be worth while to examine as to how the ratio and regression estimators in controlled simple random sampling compare with those under simple random sampling. In this note it will be shown that the ratio and regression estimators in controlled simple random sampling are *at least* as efficient as those in simple random sampling.

2. Let D'_n be an admissible (*c f* [1]) controlled simple random sampling design for drawing a sample of size $n(\geqslant 3)$ from the population $\pi : \{u_r\}$, $r=1, 2, \ldots\ldots N$, whose measure of control is α, and D_n be the simple random sampling design, so that for samples from D'_n or D_n we have

$$(2\cdot1) \quad E(\beta_r) = \frac{n}{N} \text{ and } E(\beta_r \, \beta_{r'}) = \frac{n(n-1)}{N(N-1)} \text{ for } r<r',$$

where

$$\beta_r = 1 \text{ if } u_r \text{ is included in a sample,}$$
$$= 0 \text{ otherwise.}$$

Let Y be the characteristic under study and X the ancillary character about which information is available for all u_r, $r=1,2,...,N$. Further we may assume without loss of generality that the correlation between X and Y is very high so that the regression of Y on X may be assumed to be linear.

Let $\{\bar{y}_n, \bar{x}_n\}$, $\{\bar{y}'_n, \bar{x}'_n\}$ denote the sample means corresponding to samples drawn from D_n and D'_n respectively.

Then, in virtue of Theorum (3·1) of [1], we have

$$(2\cdot2) \quad E(\bar{y}_n) = E(\bar{y}'_n) = \bar{y}_N$$

$$E(\bar{x}_n) = E(\bar{x}'_n) = \bar{x}_N$$

Also, we have

$$(2\cdot3) \quad V(\bar{y}_n) = V(\bar{y}'_n) = \frac{N-n}{Nn} S^2_y$$

$$\text{and} \quad V(\bar{x}_n) = V(\bar{x}'_n) = \frac{N-n}{Nn} S^2_x$$

where

$$(N-1)\, S^2_y = \sum_{r=1}^{N} (y_r - \bar{y}_N)^2$$

$$(N-1) S^2_x = \sum_{r=1}^{N} (x_r - \bar{x}_N)^2$$

Now consider the ratio and regression estimators for $N\bar{y}_N$ corresponding to samples drawn from D_n, namely,

$$(2\cdot4) \quad \gamma_r = N(\bar{y}_n | \bar{x}_n)\, \bar{x}_N$$

$$\text{and} \quad \gamma_L = N(\bar{y}_n + \hat{\beta}\, (\bar{x}_N - \bar{x}_n))$$

where $\hat{\beta}$ is the estimated regression co-efficient.

Under the assumption that the bias of these estimators is negligible, their mean square errors can be easily worked out and put in the following form :

$$(2.5)\text{M S.E}(\gamma_R) \cong (N\bar{y}_N)^2 \left[\frac{V(\bar{y}_n)}{\bar{y}^2_N} + \frac{V(\bar{x}_n)}{\bar{x}^2_N} - \frac{2\text{cov}(\bar{y}_n, \bar{x}_n)}{\bar{y}_N\ \bar{x}_N} \right]$$

$$\text{and} \quad \text{M.S.E}(\gamma_L) \cong N^2 V(\bar{y}_n)(1-\rho^2)$$

where 'ρ' is the coefficient of correlation between Y and X. Let us now consider the corresponding estimates in the case of samples drawn from D'_n and set

$$(2.6) \quad \gamma'_R = N(\bar{y}'_n|\bar{x}'_n)\bar{x}_N$$

$$\gamma'_L = N[\bar{y}'_n + \hat{\beta}'(\bar{x}_N - \bar{x}'_n)]$$

where $\hat{\beta}'$ is the regression coefficient estimated from sample observations.

It is easy to verify now, in virtue of (2·3), (2·4) and (2·5) that on the assumption that the bias is negligible, the mean square errors of the estimators in (2·6) differ from those given in (2·5) only in respect of the covariance between $\bar{y}'_n$ and $\bar{x}'_n$.

Now

$$(2.7) \quad \text{cov}\,(\bar{x}'_n, \bar{y}'_n) = E(\bar{x}'_n, \bar{y}'_n) - E(\bar{x}'_n)E(\bar{y}'_n)$$

$$\bar{x}^1_n\, \bar{y}^1_n = \frac{1}{n^2}\left[\sum^{n} x_i y_i + \sum^{n}_{i \neq j} x_i\, y_j\right]$$

$$= \frac{1}{n^2}\left[\sum^{N}_{r=1} \beta_r x_r y_r + \sum^{N}_{r \neq r'} \beta_r \beta_{r'}' x_r y_{r'}\right]$$

so that in view of (2·1), we have

$$(2{\cdot}8) \quad E(\bar{x}'_n\, \bar{y}'_n) = \frac{1}{n^2}\left[\frac{n}{N}\sum^{N}_{r=1} x_r y_r + \frac{n(n-1)}{N(N-1)}\sum^{N}_{r \neq r'} x_r y_r'\right]$$

Substituting from (2·8) in (2·7) and simplifying we get

$$(2{\cdot}9) \quad \text{cov}\,(\bar{x}'_n, \bar{y}'_n) = \frac{N-n}{Nn} Sxy = \text{cov}\,(\bar{y}_n, \bar{x}_n)$$

where $\quad (N-1)S_{xy} = \sum^{N}_{r=1} (x_r - \bar{x}_N)(y_r - \bar{y}_N)$

Thus from (2·9) it follows that the *M.S.E's* of the estimators given by (2·6) and those given by (2.5) are the same.

Therefore D'_n is more efficient than D_n if and only if the expected cost of survey of a sample drawn from D_n' is less than that drawn from D_n, and this, in view of Theo. (3·2) of [1], holds true when

$$0 \leqslant \alpha < 1 - \frac{\binom{N}{2}}{\binom{N}{n}}$$

Thus we get the following—

Theorem. The ratio and regression methods of estimation in controlled simple random sampling always provide more efficient estimators for $N\bar{y}_N$ than in simple random sampling when the measure α of control space satisfies the inequality,

$$0 \leqslant \alpha < 1 - \frac{\binom{N}{2}}{\binom{N}{n}}$$

which is not a very stringent restriction since α should be as small as possible.

ACKNOWLEDGEMENT

The authors are extremely thankful to Dr. V.G. Panse, Statistical Adviser, ICAR, for his constant encouragement and keen interest in this work.

REFERENCE

[1] Avadhani, M.S. and Sukhatme, B.V. "Controlled Simple Random Sampling." Jour. Ind. Soc. Agr. Stat. 1965, Vol. XVII, No. 1, pp. 34-42.