Jour. Ind. Soc. Ag. Statistics
Vol. XL, No. 3 (1988), pp. 191-201

COMPARISON OF SOME SAMPLING STRATEGIES IN THE PRESENCE OF A TREND

M. C. AGRAWAL and NIRMAL JAIN
University of Delhi, Delhi-110 007

(Received : June, 1986)

SUMMARY

It has been shown that the centred systematic sampling provides an efficiency robust estimator in the sense that it maintains its superiority vis-a-vis the other competing estimators in the face of a variety of conditions that are assumed to characterise the population. A method of sampling called 'extreme point sampling' has been proposed for populations consisting solely of linear or quadratic trend. Also modifications of Yates' and corrections and centred systematic sampling have been suggested to remove the effect of a parabolic trend. Appealing to the work of Royall and Herson [4], we have shown that the centred systematic sampling provides the best linear unbiased estimator under the superpopulation linear and quadratic trend models.

Keywords : Simple random sampling, Systematic sampling, Linear trend, Stratified sampling, Parabolic trend, Quadratic trend, BLUE.

Introduction

Consider a finite population of size N, the units of which are identified by the labels 1, 2, . . . , N and ordered in increasing size of the label. It is assumed that the population size N is expressible as a product of the sample size n and some integer k, i.e., $N = nk$. The usual systematic design classifies N units of the population into $k \geq 2$ classes $S_1, S_2, \ldots, S_k$ where S_i is then randomly selected. The population mean $\bar{Y}_N = (y_1 + y_2 + \ldots + y_N)/N$ is estimated by $\bar{y}_{si}$ if S_i is selected, where $\bar{y}_{si}$ is the mean of y-values for the n-units in S_i. The end correction method due

to Yates [5] involves the usual systematic sampling but the estimator of $\bar{Y}_N$ is

$$\bar{y}_{sy}^{(1)} = \bar{y}_{sy} + \frac{(2i - k - 1)}{2(n-1)k}(y_i - y_{i+\overline{n-1}k})$$

if S_i is selected. Method of centred systematic sampling proposed by Madow [3] is to select the class $S_{(k+1)/2}$ if k is odd and to randomly select one of the two classes $S_{k/2}$ and $S_{(k+2)/2}$, the probability of selection for each being $\frac{1}{2}$ if k is even. The estimator of $\bar{Y}_N$ is equal to $\bar{y}_{(k+1)/2}$ for k odd, to $\bar{y}_{k/2}$ or to $\bar{y}_{(k+2)/2}$ for k even.

2. Population with Linear Trend

A comparison uf the usual estimator under systematic sampling with the estimators under simple and stratified random sampling is available in the literature for a population represented by a linear trend model, with and without end corrections applied to systematic sampling. If, in place of such a linear trend model under fixed population approach, we employ a superpopulation linear trend model, the ranking of the three estimators, in order of precision remains unchanged provided the end corrections are not invoked for systematic sampling (see, e.g. Konijn, [2]). But, a natural question arises as to what will be their relative performance after the end corrections are applied to the estimator under systematic sampling of alternatively, some other method of systematic sampling is used to eliminate the linear effect.

Bellhouse and Rao [1] have compared the performances of the four methods of systematic sampling (viz., the end correction method, centred systematic sampling, modified systematic sampling and balanced systematic sampling) which eliminate the effect of the linear trend in the population. They noted that the four methods performed almost similarly under a superpopulation linear trend model which differs from our model (2.1) in the sense that they have considered $E(e_i^2) = \sigma^2$, which corresponds to $g = 0$ in (2.1). However, the authors showed that, for a superpopulation quadratic trend model (comparable to our model (3.7) with $g = 0$), the centred systematic sampling for k odd is the best followed by the systematic sampling with end corrections as the next best.

In the present and the immediately following sections, a study is undertaken with a view to comparing strategies involving systematic sampling (with and without end corrections) and centred systematic sampling (for k odd or even) with those employing simple and stratified random sampl-

ing for more general models representing linear and quadratic trends under the fixed population and superpopulation approaches.

In the context of the superpopulation approach, it would be apt to point out that a more plausible viewpoint is to assume that the finite population itself is a sample from a hypothetical infinite population called superpopulation, and the values of the successive units of the population increase in accordance with the following linear model

$$y_i = \mu + i\theta_1 + e_i \qquad (i = 1, 2, \ldots, N) \tag{2.1}$$

where μ and θ_1 are constants and the random error e_i has the properties

$$E(e_i) = 0,\ E(e_i e_j) = 0\ (i \neq j) \text{ and } E(e_i^2) = \lambda_i g\ (g \geqslant 0) \tag{2.2}$$

The symbol E denotes the expectation with respect to the superpopulation.

Let $\bar{y}_{ran}$, $\bar{y}_{st}$, $\bar{y}_{sy}$, $\bar{y}_{sy}^{(1)}$ and $\bar{y}_{sy}^{(2)}$ be the estimators of $\bar{Y}_N$ under simple random sampling, stratified random sampling, systematic sampling, systematic sampling (with Yates' end corrections) and centred systematic sampling, respectively. The corresponding average (anticipated) variances of the estimators under the above superpopulation set-up are given by

$$EV(\bar{y}_{ran}) = \frac{(k-1)(nk+1)}{12}\theta_1^2 + \frac{\lambda(k-1)}{n^2k^2}\sum_{i=1}^{N} i^g$$

$$EV(\bar{y}_{st}) = \frac{(k^2-1)}{12n}\theta_1^2 + \frac{\lambda(k-1)}{n^2k^2}\sum_{i=1}^{N} i^g$$

$$EV(\bar{y}_{sy}) = \frac{(k^2-1)}{12}\theta_1^2 + \frac{\lambda(k-1)}{n^2k^2}\sum_{i=1}^{N} i^g$$

$$EV(\bar{y}_{sy}^{(1)}) = \frac{\lambda(k-1)}{n^2k^2}\sum_{i=1}^{N} i^g + \frac{\lambda}{4(n-1)^2k^3}\sum_{i=1}^{k}(2i-k-1)^2 \{i^g + (i+\overline{n-1}k)^g\} + \frac{\lambda k-1)}{n(n-1)k^3}\sum_{i=1}^{k}(2i-k-1) \{i^g - (i+\overline{n-1}k)^g\}$$

and

$$EV(\bar{y}_{sy}^{(2)}) = \begin{cases} \dfrac{\lambda(k-2)}{n^2k}\displaystyle\sum_{i=1}^{n}\left\{\frac{(2i-1)k+1}{2}\right\}^g + \frac{\lambda}{n^2k^2}\sum_{i=1}^{N} i^g, \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{if } k \text{ is odd} \\ \dfrac{\lambda(k-2)}{2n^2k}\left[\displaystyle\sum_{i=1}^{n}\left\{\frac{(2i-1)k}{2}\right\}^g + \sum_{i=1}^{n}\left\{\frac{(2i-1)k+2}{2}\right\}^g\right] \\ \qquad + \dfrac{\lambda}{n^2k^2}\displaystyle\sum_{i=1}^{N} i^g, \quad \text{if } k \text{ is even} \end{cases}$$

The five strategies can now be compared for any value of g. On the basis of these average variances, the following conclusions emerge :

(i) The estimator $\bar{y}_{sy}^{(2)}$ is most efficient among all the estimators considered here for $g = 0, 1, 2$ and whatever may be k, even or odd.

(ii) $\bar{y}_{st}$ is superior to $\bar{y}_{ran}$, but $\bar{y}_{sy}^{(1)}$, unlike in the case of (fixed-population) linear trend model, is conditionally superior to $\bar{y}_{ran}$ and $\bar{y}_{st}$ as detailed hereafter. The estimator $\bar{y}_{sy}^{(1)}$ will perform better than the estimator $\bar{y}_{ran}$ if :

(i) $\lambda \leqslant \dfrac{(nk+1)(n-1)^2k^2}{2(k+1)}\,\theta_1^2$ for $g = 0$

(ii) $\lambda \leqslant \dfrac{(n-1)^2k^2}{(k+1)}\,\theta_1^2$ for $g = 1$

(iii) $\lambda < \dfrac{(nk+1)}{\psi(k+1)}\,\theta_1^2$ for $g = 2$

and it would be superior to $\bar{y}_{st}$ if :

(i) $\lambda < \dfrac{k^2(n-1)^2}{2n}\,\theta_1^2$ for $g = 0$

(ii) $\lambda < \dfrac{k^2(n-1)^2}{n(nk+1)}\,\theta_1^2$ for $g = 1$

(iii) $\lambda < \dfrac{\theta_1^2}{n\psi}$ for $g = 2$

where $\psi = \dfrac{(4k^2+5k-1)}{5(n-1)^2k^2} + \dfrac{(nk+1)}{(n-1)k} - \dfrac{4(k-1)}{nk}$.

3. Population with Parabolic Trend

In this section, we consider a model consisting of a quadratic component for both the fixed-population and superpopulation approaches. We shall first examine the performance of the aforesaid strategies assuming the model

$$y_i = \mu + i\theta_1 + i^2\theta_2 \qquad (i = 1, 2, \ldots, N) \tag{3.1}$$

under the fixed-population approach. μ, θ_1 and θ_2 are constants. The variances for $\bar{y}_{ran}$, $\bar{y}_{st}$, $\bar{y}_{sy}$, $\bar{y}_{sy}^{(1)}$ and $\bar{y}_{sy}^{(2)}$ under the model (3.1) are obtained as

$$V(\bar{y}_{ran}) = \frac{(k-1)(nk+1)}{12}\left[(\theta_1 + \overline{nk+1}\,\theta_2)^2 + \frac{(n^2k^2-4)}{15}\theta_2^2\right] = A \text{ (say)} \tag{3.2}$$

$$V(\bar{y}_{st}) = \frac{(k^2-1)}{12n}\left[(\theta_1 + \overline{nk+1}\,\theta_2)^2 + \frac{(5n^2k^2-4-4k^2)}{15}\theta_2^2\right] = B \text{ (say)} \tag{3.3}$$

$$V(\bar{y}_{sy}) = \frac{(k^2-1)}{12}\left[(\theta_1 + \overline{nk+1}\,\theta_2)^2 + \frac{(k^2-4)}{15}\theta_2^2\right] = C \text{ (say)} \tag{3.4}$$

$$V(\bar{y}_{sy}^{(1)}) = \frac{(k^2-1)(2k^2-3)}{60}\theta_2^2 \tag{3.5}$$

$$\text{and } V(\bar{y}_{sy}^{(2)}) = \begin{cases} \dfrac{(k^2-1)^2}{144}\theta_2^2, & \text{if } k \text{ is odd} \\ \dfrac{(k^2-4)^2}{144}\theta_2^2, & \text{if } k \text{ is even} \end{cases} \tag{3.6}$$

It is obvious that if we exclude the case $\theta_1 = -(nk+1)\theta_2$, $\bar{y}_{sy}^{(2)}$ is best of all the competing estimators considered here, followed by $\bar{y}_{sy}^{(1)}$ as the next best estimator. It may be noted that for $n = 1$, (3.2), (3.3) and (3.4) are equal. For $n > 2$ and barring the case $\theta_1 = -(nk+1)\theta_2$, the ranking of the three estimators, namely, $\bar{y}_{ran}$, $\bar{y}_{st}$ and $\bar{y}_{sy}$ is the same as in the case of a population with linear trend.

If we assume the finite population to be a sample from a superpopulation, the quadratic trend model would be written as

$$y_i = \mu + i\theta_1 + i^2\theta_2 + e_i \qquad (i = 1, 2, \ldots, N) \tag{3.7}$$

where the random error e_i has the properties as stated in (2.2).

Subject to the model (3.7), the average variances under the five sampling strategies are obtained as

$$EV(\bar{y}_{ran}) = A + \frac{\lambda(k-1)}{n^2k^2}\sum_{i=1}^{N} i^g$$

$$EV(\bar{y}_{st}) = B + \frac{\lambda(k-1)}{n^2k^2}\sum_{i=1}^{N} i^g$$

$$EV(\bar{y}_{sy}) = C + \frac{\lambda(k-1)}{n^2k^2}\sum_{i=1}^{N} i^g$$

$$EV(\bar{y}_{sy}^{(1)}) = \frac{(k^2-1)(2k^2-3)}{60}\theta_2^2 + \frac{\lambda(k-1)}{n^2k^2}\sum_{i=1}^{N} i^g + \frac{\lambda}{4(n-1)^2k^3}\sum_{i=1}^{k}(2i-k-1)^2\{i^g + (i+\overline{n-1}k)^g\} + \frac{\lambda(k-1)}{n(n-1)k^3}\sum_{i=1}^{k}(2i-k-1)\{i^g - (i+\overline{n-1}k)^g\}$$

and

$$EV(\bar{y}_{sy}^{(2)}) = \begin{cases} \dfrac{(k^2-1)^2}{144}\theta_2^2 + \dfrac{\lambda(k-2)}{n^2k^2}\displaystyle\sum_{i=1}^{N}\left\{\frac{(2i-1)k+1}{2}\right\}^g + \frac{\lambda}{n^2k^2}\sum_{i=1}^{N} i^g, & \text{if } k \text{ is odd} \\ \dfrac{(k^2-4)^2}{144}\theta_2^2 + \dfrac{\lambda(k-2)}{n^2k^2}\left[\displaystyle\sum_{i=1}^{n}\left\{\frac{(2i-1)k}{2}\right\}^g + \sum_{i=1}^{n}\left\{\frac{(2i-1)k+2}{2}\right\}^g\right] + \dfrac{\lambda}{n^2k^2}\displaystyle\sum_{i=1}^{N} i^g, & \text{if } k \text{ is even.} \end{cases}$$

It is clear from the average variances that, if we exclude the case $\theta_1 = -(nk+1)\theta_2$, the estimator $\bar{y}_{sy}^{(2)}$ retains its bestness among the five esti-

mators under superpopulation quadratic trend model for $g = 0, 1, 2$. $\bar{y}_{st}$ is superior to $\bar{y}_{ran}$ and $\bar{y}_{sy}$ excepting in the case $\theta_1 = -(nk + 1)\theta_2$. However, $\bar{y}_{sy}^{(1)}$ is more efficient than $\bar{y}_{ran}$ if :

(i) $\lambda < \dfrac{(\alpha - \xi)(n-1)^2k^2}{(k+1)}$ for $g = 0$

(ii) $\lambda < \dfrac{2(\alpha - \xi)(n-1)^2k^2}{(k+1)(nk+1)}$ for $g = 1$

(iii) $\lambda < \dfrac{2(\alpha - \xi)}{\phi(k+1)}$ for $g = 2$

and it would be superior to $\bar{y}_{st}$ if :

(i) $\lambda < (\beta - \xi')(n-1)^2k^2$ for $g = 0$

(ii) $\lambda < \dfrac{2(\beta - \xi')(n-1)^2k^2}{(nk+1)}$ for $g = 1$

(iii) $\lambda < \dfrac{2(\beta - \xi')}{\phi}$ for $g = 2$

where

$$\alpha = \frac{(nk+1)}{2}\left[(\theta_1 + \overline{nk+1}\,\theta_2)^2 + \frac{(n^2k^2-4)}{15}\theta_2^2\right]$$

$$\xi = \frac{(k+1)(2k^2-3)}{10}\theta_2^2$$

$$\beta = \frac{1}{2n}\left[(\theta_1 + \overline{nk+1}\,\theta_2)^2 + \frac{(5n^2k^2-4-4k^2)}{15}\theta_2^2\right]$$

$$\xi' = \frac{(2k^2-3)}{10}\theta_2^2$$

and

$$\phi = \frac{(4k^2+5k-1)}{5(n-1)^2k^2} + \frac{(nk+1)}{(n-1)k} - \frac{4(k-1)}{nk}$$

It is evident from the above results that, if Yates' end corrections are applied, the systematic sampling strategy performs better than simple and stratified random sampling strategies insofar as linear and quadratic

trend models under the fixed-population approach are concerned, while the same could not be said to hold under the superpopulation linear and quadratic trend models unless certain conditions spelt out earlier are satisfied.

It may be pointed out that the centred systematic sampling is devoid of the process of randomisation for k odd, and is nominally random for k even. This motivates us to present it in an entirely model-oriented perspective in Section 6.

The above discussion clearly reflects that, in considering the model (3.7), we have visualised two kinds of departures from the superpopulation model (2.1). The first kind of departure from (2.1) is conceived in the from of a quadratic component as incorporated in (3.7), and the second one relates to a breakdown of the variance function as envisaged via the choice of g ($= 0, 1, 2$). Viewed against this backdrop, the centred systematic sampling provides an efficiency robust estimator against these two kinds of departures from the superpopulation linear trend model, in the sense that it maintains its bestness (as measured by average variances) amongst the potentially conceivable competing estimators.

The notable performance of centred systematic sampling in the presence of a quadratic trend in the population can be traced to the fact that the sample selected leads to an estimate which is very close to the population value. In other words, if there is a population with a quadratic trend, the centred systematic sampling tends to substantially eliminate the effect of the quadratic trend.

4. A New Sampling Scheme

If there is a population consisting solely of a linear trend, we may adopt a simple scheme of sampling which consists in drawing the first and the last units of the population. An advantage of this sampling is that the sample mean will coincide with the population mean for both the cases $N = nk$ and $N \neq nk$. Besides, one can manage with a sample of size 2 only, whereas in other sampling schemes, we have usually $n \geqslant 2$. Such a scheme of sampling which helps to eliminate the effect of a linear trend economically will be called extreme-point sampling.

If $\bar{y}_{sy}^{(3)}$ denotes the estimator of $\bar{y}_N$ for extreme-point sampling, then $V(\bar{y}_{sy}^{(3)}) = 0$ when there is a perfect linear trend However, for the superpopulation linear trend model (2.1), the extreme point sampling strategy is as good as centred systematic sampling for $n = 2$ and $g = 0, 1$. It is thus clear that the former would also perform better than systematic sampling with and without end corrections under the conditions men-

tioned in the preceding line.

In the case of a population with a quadratic trend represented by (3.1), the y-values on the first and the last units under the extreme-point sampling may be multiplied by suitable weights in order to ensure the equality of the sample mean and the population mean, eliminating thereby the effect of a quadratic trend. The desired weights for the case $N = nk$ are worked out to be $(1 + \phi)$ and $(1 - \phi)$ for the first and the last observations, respectively, before taking the average where

$$\phi = \frac{(nk - 2)\theta_2}{3[\theta_1 + nk + 1\theta_2]}$$

5. Extended End Corrections

The extension of Yates' end corrections in the case of a population with quadratic trend modelled as (3.1) would yield the following weights by which the first and the last observations in the systematic sample (the ith column, $i' = 1, 2, \ldots, k$) should be multiplied before taking the average :

$$\text{weight for the first observation} = 1 + \eta$$

$$\text{weight for the last observation} = 1 - \eta$$

$$\text{where } \eta = \frac{n\left[\left(i' - \frac{k+1}{2}\right)\theta_1 + \left\{i'^2 + i'k(n-1) + \frac{(k+1)(k-1-3nk)}{6}\right\}\theta_2\right]}{(n-1)\,k\,[\theta_1 + \{(n-1)\,k + 2i'\}\,\theta_2]} \qquad (5.1)$$

If the above end corrections are applied to remove the effect of a parabolic trend, the systematic sample mean for the model (3.1) would reduce to

$$\bar{y}^*_{sy} = \mu + \frac{nk+1}{2}\theta_1 + \frac{(nk+1)(2nk+1)}{6}\theta_2$$

which is the population mean and hence $V(\bar{y}^*_{sy}) = 0$.

In order to eliminate the effect of a quadratic trend, the centred systematic sampling may also be subjected to end corrections. When k is odd, the weights by which the first and the last observations of the class $S(k+1)/2$ are to be multiplied before taking the average are $(1 + Z)$ and $(1 - Z)$, respectively, where Z is the value of η in (5.1) with $i' =$

$(k+1)/2$. Similarly one may easily work out the end corrections when k is even.

6. Best Linear (Model) Unbiased Estimator in the Presence of a Trend

We have remarked in Section 5 that the centred systematic sampling. is, by and large, devoid of process of randomisation. This feature motivates us to investigate the performance of the estimator based on centred systematic sampling in the light of the characteristics of the sample at hand. Hence, we will now examine such an estimator under a purely model-based approach. In what follows, the design properties are completely dispensed with.

Let x_i denote the value of the auxiliary variate on the ith unit. If we now visualise the label i itself as the auxiliary variate x_i, then the models (2.1) and (3.1) are comparable respectively to the usual linear and quadratic regression functions. Invoking the work of Royall and Herson [4], we arrive at the following conclusions :

(a) Consider the model

$$y_i = \mu + \theta_1 x_i + e_i, \qquad (\mu,\ \theta_1 \neq 0,\ i = 1, 2, \ldots, N) \qquad (6.1)$$

where $E(e_i) = 0$, $E(e_i e_j) = 0$, $E(e_i^2) = \lambda x_i$. The ratio estimator is the best linear unbiased estimator (BLUE) under the model (6.1) provided we have a balanced sample of degree 1, i.e., the sample mean of x_i's should coincide with the population mean of x_i's. Since we are considering x_i as i, the use of centred systematic sampling readily yields a balanced sample (of degree 1) when k is odd and an approximately balanced sample if k is even. Thus, for such a balanced sample, the ratio estimator would obviously reduce to an estimator which is just the mean of y-values in this sample obtained by centred systematic sampling. To summarise it, we may assert that the centred systematic sampling leads to BLUE under the model (6.1) when the label is treated as an auxiliary variate. Further, this estimator will remain optimal even when $E(e_i^2) = \lambda$.

(b) For the model

$$y_i = \mu + \theta_1 x_i + \theta_2 x_i^2 + e_i \quad (\mu, \theta_1, \theta_2 \neq 0,\ i = 1, 2, \ldots, N) \qquad (6.2)$$

where $E(e_i) = 0$, $E(e_i e_j) = 0$, $E(e_i^2) = \lambda x_i$, the usual ratio estimator is BLUE provided the following conditions are satisfied :

(i) $\frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{N} \sum_{i=1}^{N} x_i$

$$\text{(ii)}\ \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2,$$

i e., the sample is balanced up to degree 2. Since we are envisaging the auxiliary variate x_i as the label i, the centred systematic sampling furnishes a sample for which the conditions (i) and (ii) are either fully or approximately met, and hence the ratio estimator would reduce to the estimator which is mean of y values in the sample drawn with centred systematic sampling. In nutshell, the centred systematic sampling yields BLUE (or nearly BLUE) under the model (6.2). Further, the optimality of this estimator would not be disturbed even when the variance function is altered to $E(e_i^2) = \lambda$ or λ_i^2.

Besides, the optimality of the estimator based on centred systematic sampling, in the context of model (6 2), reflects that the estimator is robust against both the breakdowns in the model (6.1) consisting of a quadratic term and the alteration in the variance function as spelt out above.

REFERENCES

[1] Bellhouse, D. R. and Rao, J. N. K. (1975) : Systematic sampling in the presence of a trend, *Biometrika* **62** : 694-697.

[2] Konijn, H. S. (1973) : *Statistical Theory of Sample Survey Design and Analysis*, North Holland Publishing Company, Amsterdam.

[3] Madow, W. G. (1953) : On the theory of systematic sampling, III, *Ann. Math. Stat.* **24** : 101-106.

[4] Royall, R. M. and Herson, J. (1973) : Robust estimation in finite populations I, *JASA* **68** : 880-889.

[5] Yates, F. (1948) : Systematic sampling, *Phil. Trans. Roy. Soc., London*, **A241** : 345-377.