Jour. Ind. Soc. Ag. Statistics
Vol. XXXVII, No. 2 (1985), pp. 121-130

ON THE EFFICIENCY OF THE RATIO ESTIMATORS UNDER SIZE STRATIFICATION

PRANESH KUMAR
Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi 110 012

(Received : January, 1981)

SUMMARY

The efficiency of the ratio estimator under size stratification has been worked out depending upon the size character when it is discrete and has also been compared with some of the well-known sampling strategies. On comparing the efficiencies, it has been established that the stratified ratio sampling strategy under reference performs satisfactorily.

Keywords : Ratio Estimator; Size Stratification; Unbiased Regression Estimator.

1. Introduction

When information on an ancillary character, say x, which is highly correlated with the character of interest (y) is available, it may be used for providing more accurate estimate of the population parameter in question either by selecting the sample by simple random sampling and using ratio or regression method of estimation, or by selecting the units with probabilities proportional to their X values. Reference in this direction may be made to the papers of Hansen and Hurwitz [8], Madow [13], Narain [14], Hartley and Rao [11], Rao, Hartley and Cochran [15], Hanurav [9], [10], Fellegi [6], Hajek [7], Durbin [5], Vijayan [18], Singh and Srivastava [16] etc. Alternatively, the ancillary information may be used for stratifying the survey population. Following the pioneering work of Dalenius [4], reference may be made to the comprehensive works of Cochran [3] and Singh [17]. Thus, though there exists a fairly good number of techniques for utilisation of the ancillary information for building up better estimates of population parameters, none of them is entirely satisfactory. Either they are based on unrealistic assumptions or the solutions are so complicated that they can be of very little practical use or they are too approximate. Avadhani [1] suggested an estimator under size stratification depending on the ancillary information whether discrete or continuous. In this paper an attempt has been made to examine the possibility of using the classical ratio estimator under size stratification suggested by Avadhani [1] and comparing the resulting strategy with some of the *PPS* sampling strategies.

2. Notations

Let the population under study consist of N distinct units $i = 1, 2, \ldots, N$. Set y to denote the character of interest and x the ancillary character which is highly correlated with y. Let Y_i and X_i be, respectively, y and x character values of the ith unit where it is assumed throughout in what follows that $X_i > 0$ and is known for all $i = 1, 2, \ldots, N$.

In sample survey situations generally the population total

$$Y = \sum_{i=1}^{N} Y_i$$

or, the population mean

$$\bar{Y} = Y/N$$

is the parameter of interest to be estimated with the help of a sample of n observations.

3. Stratified Ratio Sampling Strategy Using Discrete Ancillary Information

In practice, the x-character may be discrete in the sense that there exist only k-distinct x-values, say $X_1, X_2 \ldots X_k$, each occurring $N_1, N_2 \ldots N_k$ times where $N_1 + N_2 + \ldots + N_k = N$.

Let y and x be such that the corresponding values of the units satisfy the following structure :

$$\left.\begin{array}{l} Y_{ij} = \beta X_i + e_{ij}, \qquad j = 1, 2, \ldots, N_i \\ \qquad\qquad\qquad\qquad \text{and } i = 1, 2, \ldots, k \\ \text{with } \sum_{j=1}^{N_i} e_{ij} = 0, \\ \text{and } S^2_{ie} = (N_i - 1)^{-1} \sum_{j=1}^{N_i} e^2_{ij} = \gamma X_i^g, \\ \text{where } g \geqslant 0 \text{ and } \gamma > 0. \end{array}\right\} \qquad (3.1)$$

Without loss of generality, we assume that $k < n$ as otherwise it can be ensured by suitably pooling neighbouring x's. However, in this case, properties of the estimator such as unbiasedness, variance will be affected. Thus, the suggested scheme is applicable only to the situations wherein $k < n$.

3.1 *Sampling Procedure 'A_1'*

The sampling procedure proposed by Avadhani [1] when model (3.1) holds is as follows :

A_1 (*I*) Treat the k groups with $N_1, N_2, \ldots, N_k$ units where $\sum_{i=1}^{k} N_i = N$, as strata.

A_1 (*II*) Select independent samples of sizes $n_1, n_2, \ldots, n_k$ where $\sum_{i=1}^{k} n_i = n$, from these k strata by simple random sampling without replacement.

Suppose a sample of size n units is drawn from the population by this sampling procedure (A_1). Let $\bar{y}_i$ denote the sample mean of the n_i observations from the ith stratum, $i = 1, 2, \ldots, k$. Set $W_i = N_i/N$ and $\bar{y}_{ST} = \sum_{i=1}^{k} W_i \bar{y}_i$ which is the classical unbiased stratified sample estimator of the population mean, $\bar{Y}$. The ancillary information may be utilised for improving this estimator by the ratio/regression method of estimation. If $\bar{y}_{CR}$ and $\bar{y}_{SR}$ denote, respectively, the combined ratio and the separate ratio estimators, we have the following :

THEOREM 1. *Under the sampling procedure 'A_1',*

(i) $\bar{y}_{CR} = \bar{y}_{SR} = \bar{y}_{ST}$,

$$\text{(ii)}\ V(\bar{y}_{CR}) = \frac{\Upsilon}{n}\left(\sum_{i=1}^{k} W_i X_i^{g/2}\right)^2 - \frac{\Upsilon}{n}\sum_{i=1}^{k} W_i X_i^{g}, \qquad (3.2)$$

when $$n_i = nN_i X_i^{g/2}\left(\sum_{i=1}^{k} N_i X_i^{g/2}\right)^{-1}$$

and

$$\text{(iii)}\ V(\bar{y}_{CR}) = \Upsilon\left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} W_i X_i^{g}, \qquad (3.3)$$

when $n_i = n\, W_i$.

Proof

(i) By virtue of (3.1), $\beta = R$ and therefore, the estimators $\bar{y}_{CR}$ and $\bar{y}_{SR}$ are equally efficient as the corresponding regression estimators. Also since $\bar{X}_i = X_i$, $i = 1, 2, \ldots, k$, it could be easily established that

$$\bar{y}_{CR} = \bar{y}_{SR} = \bar{y}_{ST} \tag{3.4}$$

(ii) and (iii). It is seen from (3.4) that

$$V(\bar{y}_{CR}) = \sum_{i=1}^{k} W_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_{iy}^2.$$

By virtue of (3.1), $V(\bar{y}_{CR})$ becomes

$$V(\bar{y}_{CR}) = \Upsilon \sum_{i=1}^{k} W_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) X_i^g \tag{3.5}$$

Now (3.2) or (3.3) follows from (3.5) according as

$$n_i = n\, N_i X_i^{g/2} \left(\sum_{i=1}^{k} N_i X_i^{g/2} \right)^{-1}$$

or $n_i = n\, W_i$.

The sampling procedure 'A_1' together with $\bar{y}_{CR}$ will hereafter be denoted by $S(A^1_{CR})$ and will be known as stratified ratio sampling strategy.

4. Comparison of Efficiencies

In this section the sampling strategy $S(A^1_{CR})$ has been compared with Midzuno-Sen (1952) Ratio Sampling Strategy; *pps* sampling strategies of Hartley and Rao [11] and Rao, Hartley and Cochran [15] and sampling strategy of Singh and Srivastava [16].

It has been assumed throughout in what follows that the x and y character values of the population units satisfy the model (3.1).

If a strategy $S_{T'}$ is more efficient than $S_{T''}$, in the sense that $V(T') < V(T'')$, we write for brevity $S_{T'} > S_{T''}$ and if they are equally efficient i.e. $V(T') = V(T'')$, we designate it by $S_{T'} \sim S_{T''}$.

4.1 *Midzuno-Sen Ratio Sampling Strategy*

Since the probability of selecting a sample of n distinct units from a population of size N under the Midzuno-Sen procedure is

$$p_s = \bar{x}/\tbinom{N}{n}\,\bar{X},$$

where $\bar{x}$ denotes the mean of the x-characteristic defined on the sample units $i_1, i_2, \ldots, i_n$ and $\bar{X} = N^{-1} \sum_{i=1}^{N} X_i$, it is well known that the ratio estimator

$$\bar{y}_R = \bar{y}\,\bar{X}/\bar{x}$$

provides an unbiased estimate of the population mean $\bar{Y}$.

Avadhani and Srivastava [2] have shown that in large samples, the variance of $\bar{y}_R$ is given by

$$V(\bar{y}_R) = \sum_{i=1}^{N} (Y_i - R\,X_i)^2\,(N - n)/nN, \tag{4.1}$$

where $R = \bar{Y}/\bar{X}$, and which is exactly the same as that of ratio estimator under simple random sampling without replacement.

Midzuno-Sen sampling procedure together with the ratio estimator $\bar{y}_R$, will hereafter be referred to as $S\,(^{MS}_{R})$ strategy.

Now we have the following theorem:

THEOREM 2. *If* $n_i = n\,N_i\,X_i^{g/2}\left(\sum_{i=1}^{k} N_i X_i^{g/2}\right)^{-1}$, *then* $S\,(A^1_{CR}) > S\,(^{MS}_{R})$, *for all* $g > 0$ *except* $g = 0$ *for which* $S\,(A^1_{CR}) \sim S\,(^{MS}_{R})$, *provided of course, n is sufficiently large so thal* $\bar{x} \to \bar{X}$.

Proof. By virtue of (3.1), it is apparent that

$$V(\bar{y}_R) = \gamma\,(N - n) \sum_{i=1}^{k} W_i X_i^g/nN, \tag{4.2}$$

provided N_i and N are sufficiently large so that $(N_i - 1)/(N - 1) = W_i$.

From (3.2) and (4.2), it is seen that

$$\begin{aligned} V(\bar{y}_R) - V(\bar{y}_{CR}) &= \gamma\left[\sum_{i=1}^{k} W_i \left\{X_i^{g/2} - \sum_{i=1}^{k} W_i X_i^{g/2}\right\}^2\right]\Big/ n \qquad (4.3)\\ &> 0 \quad \text{for } g > 0\\ &= 0 \quad \text{for } g = 0 \end{aligned}$$

4.2 *Hartley and Rao Sampling Strategy*

Under the Hartley and Rao [11] sampling procedure, the Horvitz-Thompson [12] estimator

$$\bar{y}_{HT} = \sum_{i=1}^{n} Y_i\,X/nN\,X_i \tag{4.4}$$

of population mean and its variance

$$V(\bar{y}_{HT}) = \sum_{i=1}^{N} X_i \{X - (n-1) X_i\} \{(Y_i/NX_i) - \bar{Y}\}^2/n \tag{4.5}$$

will be designated as $S(^{HR}_{PPS})$ sampling strategy.The comparison of $S(A^1_{CR})$ and $S(^{HR}_{PPS})$ gives the following.

THEOREM 3. *In all finite populations wherein the model* (3.1) *holds,*

$$S(^{A_1}_{CR}) > S(^{HR}_{PPS}) \text{ for all } g \geqslant 0, \text{ if } n_i = n\, N_i\, X_i^{g/2} \left(\sum_{i=1}^{k} N_i X_i^{g/2}\right)^{-1}.$$

Proof. By virtue of (3.1) and (4.5), we have

$$V(\bar{y}_{HT}) = \left\{\gamma\, \bar{X} \sum_{i=1}^{k} W_i X_i^{g-1} \Big/ n\right\} - \left\{\gamma \sum_{i=1}^{k} W_i X_i^{g} \Big/ N\right\}$$
$$+ \gamma \sum_{i=1}^{k} W_i X_i^g/nN, \tag{4.6}$$

provided N_i and N are sufficiently large so that $(N_i - 1)/(N - 1) \doteq W_i$. Equations (3.2) and (4.6) yield into
$V(\bar{y}_{HT}) - V(\bar{y}_{CR})$

$$= \gamma \left[\left(\sum_{i=1}^{k} W_i X_i^{g-1}\right)\left(\sum_{i=1}^{k} W_i X_i\right) - \left(\sum_{i=1}^{k} W_i X_i^{g/2}\right)^2\right] \Big/ n$$
$$+ \frac{\gamma}{nN} \sum_{i=1}^{k} W_i X_i^g \tag{4.7}$$

By Cauchy-Schwarz's inequality, we have

$$\left(\sum_{i=1}^{k} W_i X_i^{g-1}\right)\left(\sum_{i=1}^{k} W_i X_i\right) \geqslant \left(\sum_{i=1}^{k} W_i X_i^{g/2}\right)^2$$

Thus,

$$V(\bar{y}_{HT}) - V(\bar{y}_{CR}) > 0 \quad \text{for all } g \geqslant 0.$$

4.3 *Rao, Hartley and Cochran Sampling Strategy*

If Y_j, $j = 1, 2, \ldots, n$ denote the y-character values of the n units selected by the Rao, Hartley and Cochran [15] procedure, the estimator of population mean is given by

$$\bar{y}_{RHC} = \sum_{j=1}^{n} Y_j\, p'_j\, X/N\, X_j, \tag{4.8}$$

where p'_j denotes the sum of the probabilities of the units falling in jth group.

The variance of $\bar{y}_{RHC}$ is shown by these authors to be minimum when $N_1 = N_2 = \ldots = N_n$ and is given by

$$V(\bar{y}_{RHC}) = X\{(N-n)/(N-1)n\} \sum_{i=1}^{N} X_i \{(Y_i/N X_i) - \overline{Y}\}^2 \qquad (4.9)$$

The estimator $\bar{y}_{RHC}$ under the selection procedure of Rao *et al.* along with $V(\bar{y}_{RHC})$ will be termed as the *RHC* sampling strategy and will be denoted by $S\left({}^{RHC}_{PPS}\right)$.

Now we have the following:

TNEOREM 4. *If* $n_i = n N_i X_i^{g/2} \left(\sum_{i=1}^{k} N_i X_i^{g/2} \right)^{-1}$, *then*

$$S\left({}^{A_1}_{CR}\right) > S\left({}^{RHC}_{PPS}\right), \text{ for all } g \geqslant 0.$$

Proof. Since Avadhani and Srivastava [2] have shown that

$$S\left({}^{MS}_{R}\right) > S\left({}^{RHC}_{PPS}\right) \quad \text{for } 0 \leqslant g < 1$$

and in Theorem 2 we have shown that

$$S\left({}^{A_1}_{CR}\right) > S\left({}^{MS}_{R}\right) \quad \text{for g} > 0,$$

it follows that

$$S\left({}^{A_1}_{CR}\right) > S\left({}^{RHC}_{PPS}\right) \quad \text{for } 0 \leqslant g < 1.$$

Now using (3.1) and simplifying (4.9), we see

$$V(\bar{y}_{RHC}) = \gamma\, \overline{X}\, (N-n) \sum_{i=1}^{k} W_i X_i^{g-1}/n N, \qquad (4.10)$$

provided N_i and N are sufficiently large such that $(N_i - 1)/(N-1) \doteq W_i$. From (3.2) and (4.10), we get

$$V(\bar{y}_{RHC}) - V(\bar{y}_{CR}) = \gamma \operatorname{Cov}(X, X^{g-1})/n + \gamma \left[\left(\sum_{i=1}^{k} W_i X_i^{g-1} \right) \left(\sum_{i=1}^{k} W_i X_i \right) - \left(\sum_{i=1}^{k} W_i X_i^{g/2} \right)^2 \right] \Big/ N \qquad (4.11)$$

Now since the second term of R.H.S. is a positive quantity as

$$\left(\sum_{i=1}^{k} W_i X_i^{g-1} \right) \left(\sum_{i=1}^{k} W_i X_i \right) > \left(\sum_{i=1}^{k} W_i X_i^{g/2} \right)^2$$

by Cauchy = Schwarz's inequality and the first term is positive for all $g \geqslant 1$, hence

$$V(\bar{y}_{RHC}) - V(\bar{y}_{RC}) > 0 \quad \text{for all } g \geqslant 1$$

Thus, $S\binom{A_1}{CR} > S\binom{RHC}{PPS}$ for all $g \geqslant 0$.

4.4 *Singh and Srivastava Sampling Strategy*

Singh and Srivastava [16] have suggested a sampling scheme for which the usual regression estimator is unbiased. They have also considered another sampling scheme with an unbiased regression type estimator. Since they have found first scheme to be more satisfactory than the other with respect to the efficiency, we will consider the first scheme for comparing the efficiency of the scheme given in Section 3.

Under the sampling scheme *I* of Singh and Srivastava [16] the probability of selecting a sample *s* is

$$p_s = s_x^2 / \binom{N}{n} S_x^2,$$

where $s_x^2 = (n-1)^{-1} \sum_{r=1}^{n} (x_r - \bar{x})^2$, $S_x^2 = (N-1)^{-1} \sum_{i=1}^{N} (x_i - \bar{X})^2$.

$$\bar{x} = n^{-1} \sum_{r=1}^{n} x_r, \quad \bar{X} = N^{-1} \sum_{i=1}^{N} x_i$$

Under this sampling scheme, all samples should have non-zero s_x^2.

Further, these authors have shown that for a large population, the variance of the estimator $\bar{y}s = \bar{y} + b(\bar{X} - \bar{x})$, where $\bar{y} = n^{-1} \sum_{r=1}^{n} y_r$ and $b = (n-1)^{-1} \sum_{r=1}^{n} y_r(x_r - \bar{x})/s_x^2$, is given by

$$V(\bar{y}s) = (n^{-1} + n^{-2})(1 - \rho^2)\,\mu_{02} + n^{-2}\,\mu_{20}^{-1}\left(\frac{2\mu_{11}\,\mu_{31}}{\mu_{20}} - \frac{\mu_{11}^2\mu_{40}}{\mu_{20}^2} - \mu_{22}\right)$$

Here, μ_{pq} is the *pq*th central moment of random variables *x* and *y*.

Singh and Srivastava sampling scheme together with the regression estimator $\bar{y}s$ will hereafter be referred to as $S\binom{SS}{REG}$ sampling strategy.

THEOREM 5. *For large populations, with* $n_i = n\,N_i\,x_i^{g/2}\left(\sum_{i=1}^{k} N_i x_i^{g/2}\right)^{-1}$

$$S\binom{1}{CR} < S\binom{SS}{REG} \qquad \text{for } g = 0$$

and

$$S\binom{A_1}{CR} > S\binom{SS}{REG} \qquad \text{for } g > 0$$

$$\textit{iff}\ \Delta_x > \gamma^{-1}\beta^2$$

Here, $\Delta_x = \sum_{i=1}^{k} W_i \left\{x_i^{g/2} - \left(\sum_{i=1}^{k} W_i\, x_i^{g/2}\right)\right\}^2.$

Proof. Simplifying (4.12), using (3.1) and neglecting terms of order n^{-2}, it is seen that

$$V(\bar{y}_S) \simeq \frac{\gamma}{n} \sum_{i=1}^{k} W_i\, x_i^{g} - \frac{\beta^2}{n} \qquad (4.13)$$

Further for large N;

$$V(\bar{y}_{CR}) \simeq \frac{\gamma}{n} \left(\sum_{i=1}^{k} W_i\, x_i^{g/2}\right)^2 \qquad (4.14)$$

A comparison of (4.13) and (4.14) yields

$$V(\bar{y}_S) - V(\bar{y}_{CR}) = \frac{\gamma}{n} \sum_{i=1}^{k} W_i \Big\{\Big(x_i^{g/2} - \Big(\sum_{i=1}^{k} W_i\, x_i^{g/2}\Big)\Big\}^2 - \frac{\beta^2}{n} \qquad (4.15)$$

From (4.15), it is obvious that

$$V(\bar{y}_S) < V(\bar{y}_{CR}), \quad \text{for } g = 0$$

and

$$V(\bar{y}^S <)\ V(\bar{y}_{CR}) \quad \text{for } g > 0,$$

iff

$$\gamma \sum_{i=1}^{k} W_i \left\{x_i^{g/2} - \left(\sum_{i=1}^{k} W_i\, x_i^{g/2}\right)\right\}^2 > \beta^2$$

This proves the theorem.

It is worth mentioning at this stage that the results obtained on the comparisons of various sampling strategies are model-based and hold when the population satisfies the finite population model. Further, these results will be identical to those obtained under its equivalent super-population model.

ACKNOWLEDGEMENT

The author is thankful to the referees whose comments improved considerably the form of the paper.

REFERENCES

[1] Avadhani, M. S. (1972): On optimum utilisation of the ancillary information in sampling from finite populations, *Proc. 59th session of the Indian Science Congress, Calcutta.*

[2] Avadhani, M. S. and Srivastava, A. K. (1972): A comparison of Midzuno-Sen scheme with pps sampling without replacement and its application to successive sampling, *Ann. Inst. Stat. Math.* **24** (1): 153-164.

[3] Cochran, W. G. (1961): Comparison of methods determining stratum boundaries, *Bull. Int. Stat. Inst.* **38**: 345-358.

[4] Dalenius, T. (1950): The problem of optimum stratification, *Skand. Akt.* **33**: 203-213.

[5] Durbin, J. (1967): Estimation of sampling error in multistage-surveys, *Appl. Stat.* **16**: 152-164.

[6] Fellegi, I. P. (1963): Sampling with varying probabilities without replacement: rotating and non-rotating samples, *Jour. Amer. Stat. Assoc.* **58**: 183-201.

[7] Hajek, J. (1964): Asymptotic theory of rejective sampling with varying probabilities from a finite population, *Ann. Stat.* **35**: 1491-1523.

[8] Hansen, M. H. and Hurwitz, W. N. (1943): On the theory of samplling from finite populations, *Ann. Math. Stat.* **14**: 333-362.

[9] Hanurav, T. V. (1962): Some sampling schemes in probability sampling, *Sankhya, A.* **24**: 227-230.

[10] Hanurav, T. V. (1967): Optimum utilization of auxiliary information: πps sampling of two units from a stratum, *Jour. Roy. Stat. Soc. B,* **29**: 374-391.

[11] Hartley, H. O. and Rao, J. N. K. (1962): Sampling with unequal probabilities and without replacement, *Ann. Math. Stat.* **33**: 350-374.

[12] Horvitz, D. G. and Thompson, D. J. (1952): A generalization of sampling without replacement from a finite universe, *Jour. Amer. Stat. Assoc.*, **47**: 663-685.

[13] Madow, W. G. (1949): On the theory of systematic sampling-II, *Ann. Math. Stat.* **20**: 333-354.

[14] Narain, R. D. (1951): On sampling without replacement with varying probabilities, *Jour. Ind. Soc. Agri. Stat.* **3**: 169-174.

[15] Rao, J. N. K., Hartley, H. O. and Cochran, W. G. (1962): On a simple procedure of unequal probability sampling without replacement, *Jour. Roy. Stat. Soc., B,* **24**: 482-491.

[16] Singh, P. and Srivastava, A. K. (1980): Sampling schemes providing unbiased regression estimators, *Biometrika* **67**: 205-209.

[17] Singh, R. (1967): Some contributions to the theory of construction of strata, *Unpublished Ph. D. thesis, I. A. R. I., New Delhi.*

[18] Vijayan, K. (1968): An exact πps sampling scheme—Generalization of a method of Hanurav, *Jour. Roy. Stat. Soc. B,* **30**: 556-566.