ON SOME PROPERTIES OF THE HORVITZ—THOMPSON ESTIMATOR BASED ON MIDZUNO'S ΠPS SAMPLING SCHEME

ARIJIT CHAUDHURI
Department of Statistics, Calcutta University
(Received : May, 1976)

INTRODUCTION

In a recent article Avadhani and Srivastava [1] have expressed the opinion that the strategy of estimating a finite population total by choosing a sample according to Midzuno-Sen [10], [14] scheme and employing the Horvitz-Thompson [9] estimator (*H T E*, in brief) based on a sample so drawn is unsatisfactory and is inferior in practice, from the point of view of efficiency, to Hansen-Hurwitz [7] and Rao-Hartley-Cochran [13] strategies. Chaudhuri [4] and Mukhopadhyay [11], however, have considered a slight modification (imposing, of course, a condition on the original normed size-measures) of the first-mentioned strategy and shown that the modified strategy yields a variance of estimate uniformly smaller than the variance of the estimate based on Hansen-Hurwitz strategy. This modification (originally suggested by Rao [12] in case of samples of size 2) is justified on the ground that the *HTE* is expected to be efficient when it is based on some Π*PS* sampling design (vide Hanurav [8]). Avadhani and Sukkhatme [2], [3] and Avadhani and Srivastava [1] considered a particular model under which they compared different strategies namely Rao-Hartley-Cochran (*RHC*, in short) strategy, ratio method of estimation based on (1) simple random sampling without replacement (SRSWOR) and (2) Midzuno-Sen scheme of sampling scheme. Hanurav [8] on the other hand considered a different model (which is rather a customary one in finite population sampling and is generally called a superpopulation model ; vide Cochran [6] to compare the efficiency of Rao-Hartley-Cochran strategy with the strategy of Horvitz-Thompson (*H T*) estimation based on any Π*PS* sampling scheme. In this article we consider both of these models and study the performances of the *HT* method of estimation based on the modified Midzuno-Sen scheme compared to other strategies and find that this strategy fares well and even better than the *RHC* strategy under situations generally met with in practice. Whenever the nature of the available size-measures permits the applicability of this strategy it deserves more attention than it is receiving now because it can be implemented more easily than *RHC* strategy, for example.

2. Notations and the results

We shall use the following notations :

$U = (1, \ldots\ldots, i, \ldots\ldots, N)$: a finite population of N identifiable units tagged with the labels $i = 1, \ldots\ldots N$;

y : a real-variate defined on U assuming the value Y_i on the i-th unit $(i=1, \ldots\ldots\ldots, N)$ with $Y = (Y_1, \ldots\ldots\ldots, Y_i, \ldots\ldots\ldots, Y_N)$ and $T = \sum_1^N Y_i$, the population total required to be estimated on the basis of a sample s of units selected from U (the sample-size being n with $2 \leqslant n < N$) ; X_i : the size-measures of the units such that $X_i > 0$, for all $i = 1, \ldots\ldots N, X = \sum_1^N X_i$, $\bar{X} = X/N$, $p_i = X_i/X$ $(i = 1, \ldots\ldots, N)$ with the assumption throughout this paper that $np_i < 1$ for all $i = 1, \ldots\ldots, N$.

We shall consider the following strategies (details are omitted for the sake of brevity and may be read from the references cited), namely,

I. Hansen-Hurwitz [7] strategy with the estimator and variance of the estimator as

$$t_1 = \left(\frac{1}{n}\right) \sum_{i \in s} Y_i f_i / p_i,$$

(f_i = frequency of i-th unit in s) and

$$V(t_1) = \frac{1}{n} \sum p_i \left(Y_i/p_i - T\right)^2;$$

II. Rao-Hartley-Cochran [13] strategy with the usual estimator denoted as t_2 with variance as

$$V(t_2) = \frac{N-n}{n(N-1)} \sum p_i \left(\frac{Y_i}{p_i} - T\right)^2;$$

III. Sampling according to Midzuno-Sen [10, 14] scheme and using the estimator $t_3 = X\left(\frac{\bar{y}}{\bar{x}}\right)$ ($\bar{y}, \bar{x}$ = sample means of Y's and X's) with its variance denoted as $V_M(t_3)$;

IV. Simple random sampling without replacement (SRSWOR) andusing the estimator t_3 with its Mean Square Error denoted as $V(t_3)$;

V. Modified Midzuno-Sen strategy as considered by Chaudhuri [4] and Mukhopadhyay [11] using the estimator

$$t_4 = \left(\frac{1}{n}\right) \sum_{i \in s} Y_i/p_i,$$

assuming

$$1 > np_i > \frac{(n-1)}{N-1} \text{, for all } i = 1, \ldots\ldots, N \qquad \ldots(2.1)$$

with variance

$$V(t_4) = \left(\frac{n-1}{N-1}\right)\frac{1}{n(N-2)}\left[\sum_i \left(\frac{Y_i}{p_i} - T\right)^2 - \left\{\sum_i \left(\frac{Y_i}{p_i} - T\right)\right\}^2\right]$$
$$+ \left(\frac{N-2n}{N-2}\right)\frac{1}{n} \sum_i p_i \left(\frac{Y_i}{p_i} - T\right)^2.$$

We shall consider the following two models considered by Avadhani and Sukhatme [3] and Hanurav [8] among others, namely,

$$M_1 : Y_i = \beta X_i + e_i$$
$$(i = 1, \ldots\ldots, N)$$

such that

$$\sum_{i=1}^{N} e_i = 0 = \sum_{i=1}^{N} e_i X_i,$$
$$\bar{e}_i^2 = \sigma^2 X_i^g,$$
$$0 < \sigma < +\infty,$$
$$0 \leqslant g \leqslant 2$$

$\bar{e}_i$ being the average of the e_i's in the array for which X_i is fixed ;

$$M_2 : Y_i = \beta X_i + e_i$$
$$(i = 1, \ldots\ldots, N)$$

with

e_i's as random variables such that

$$\epsilon(e_i) = 0 = \epsilon(e_i e_j) \text{ for all } i, j \ (i \neq j)$$
$$\epsilon(e_i^2) = \sigma^2 X_i^g,$$
$$0 < \sigma < \infty,$$
$$0 \leqslant g \leqslant 2$$

(ϵ is the expectation operator with respect to the distribution of e_i's assumed in the model M_2). Chaudhuri [4] and Mukhopadhyay [11] have established that

$V(t_4) < V(t_2)$ uniformly in Y provided (2.1) holds. Now, we note, first, that

$$V(t_2)-V(t_4) = \frac{1}{n}\sum p_i\left(\frac{Y_i}{p_i}-T\right)^2\left[\left(\frac{N-n}{N-1}-\frac{N-2n}{N-2}\right)\right.$$
$$\left.-\frac{n-1}{N-1}\cdot\frac{1}{N-2}\cdot\frac{1}{p_i}\right]$$
$$+\left(\frac{n-1}{N-1}\right)\frac{1}{n(N-2)}\left[\left\{\sum\left(\frac{Y_i}{p_i}-T\right)\right\}^2\right]$$

and hence on noting the relation (2.1) we get

Theorem 1. If N be so large that we may neglect the error in replacing $(N-1)$ by $(N-2)$, then

$V(t_2) > V(t_4)$ uniformly in Y.

Remark I. One may recall that Chaudhuri [5] obtained a sufficient condition for the variance of the *HTE* based on any πPS sampling scheme to be uniformly smaller than $V(t_2)$ but also observed that condition to be unrealizeable in practice.

Next, assuming the model M_1, we have

$$V(t_2) = \frac{N-n}{n(N-1)}X\sum\frac{e_i^2}{X_i} = \frac{N-n}{n(N-1)}\sigma^2 X\sum X_i^{g-1}$$

$$V(t_4) = \frac{N-2n}{N-2}\cdot\frac{1}{n}\cdot X\sum\frac{e_i^2}{X_i}$$
$$-\left(\frac{n-1}{N-1}\right)\frac{X^2}{n(N-2)}\sum_{i\neq j}\sum\frac{e_i\, e_j}{X_i X_j}$$
$$=\left(\frac{N-2n}{N-2}\right)\frac{1}{n}\,\sigma^2 X\sum X_i^{g-1}$$
$$-\left(\frac{n-1}{N-1}\right)\frac{X^2}{n(N-2)}\,\sigma^2\sum_{i\neq j}\sum X_i^{t-1}\,X_j^{t-1}$$

(writing $g=2t$).

Hence follows

Theorem 2. If the model M_1 holds, then

$$V(t_2) > V(t_4).$$

Next, we may recall from Avadhani and Srivastava [1] that if the model M_1 holds, then we have (i) approximately (approximation due to neglecting the error in replacing $\bar{X}$ by $\bar{x}$ for each s),

$$V_M(t_3) = V(t_3) = \frac{N-n}{N-1}\cdot\frac{N}{n}\cdot\sum e_i^2$$
$$=\frac{N-n}{N-1}\cdot\frac{N}{n}\sigma^2\sum X_i^{g},$$

and

(*ii*) $V(t_3) = V_M(t_3) \geqslant V(t_2)$ if $g \geqslant 1$

and $V(t_3) < V(t_2)$ if $0 \leqslant g < 1$

Hence we get

Theorem 3. If the model M_1 holds,

then

$$V_M(t_3) = V(t_3) > V(t_4) \quad \text{if} \quad g \geqslant 1.$$

Remark II. We cannot say anything definite about the sign of $[V(t_3) - V(t_4)]$ if $0 \leqslant g < 1$, but it is well-known that in practice $1 \leqslant g \leqslant 2$.

From Hanurav [8] it may be noted that if the model M_2 holds, then [because t_4 is based on a πPS design]

$$\epsilon[V(t_2)] > \epsilon[V(t_4)], \quad \text{if} \quad g > 1$$

and

$$\epsilon[V(t_2)] \leqslant \epsilon[V(t_4)], \quad \text{if} \quad g \leqslant 1.$$

When the model is M_2 and not M_1, $V(t_3)$ and $V_M(t_3)$ are not approximately same ; but in this case we may note that we have

$$\epsilon[V(t_3) - V(t_4)] = \frac{\sigma^2}{n}\left[\left\{\left(\frac{N-n}{N-1}\right) + \frac{n}{N}\right\} N\Sigma X_i^g - X\Sigma X_i^{g-1}\right]$$

$$> \frac{\sigma^2}{n}\left\{N\Sigma X_i^g - X\Sigma X_i^{g-1}\right\}$$

and hence follows the following

Theorem 4. If the model M_2 holds, then in respect of the ratio estimator t_3 based on SRSWOR, we have

$$\epsilon[V(t_3)] > \epsilon[V(t_4)] \quad \text{if} \quad g \geqslant 1.$$

Finally, we note that if the model M_2 holds, then assuming that the error in replacing $\bar{X}$ by $\bar{x}$ is negligible for each s, as in Avadhani and Srivastava [1], then one has

$$\epsilon\left[V_M(t_3)\right] = \sigma^2 \frac{N-n}{n} \cdot N\Sigma X_i^g$$

Hence we get

Theorem 5. If the model M_2 holds, and we neglect the error in replacing $\bar{X}$ by $\bar{x}$ for each s, then it follows that

$$\epsilon[V_M(t_3) - V(t_4)] \geqslant 0 \quad \text{if} \quad g \geqslant 1$$

and

$$\epsilon[V_M(t_3) - V(t_4)] < 0 \quad \text{if} \quad g < 1.$$

Summary

Efficiency of the sampling strategy of basing the Horvitz-Thompson estimator on Midzuno's sampling scheme suitably modified into πPS scheme is studied in relation to Rao-Hartley-Cochran strategy and ratio-estimation based on Midzuno's sampling scheme and simple random sampling schemes without replacement. Two well-known models are assumed and some approximations are also considered.

References

[1] 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, 153-164.

[2] Avadhani, M.S. and Sukhatme, B.V. (1972) Sampling on several successive occasions with equal and unequal probabilities and without replacement. *Aust. Jour. Statist.*, 14, 109-119.

[3] Avadhani, M.S. and Sukhatme, B.V. (1970) A comparison of two sampling procedures with an application to successive sampling. *Jour. Roy. Statist. Soc.* (C). *Applied Statistics*, 19, 251-259.

[4] Chaudhuri, Arijit (1974) On some properties of the sampling scheme due to Midzuno. *Cal. Stat. Assoc. Bull.*, 23, 1-19.

[5] ,, ,, (1975) Some properties of estimators based on sampling schemes with varying probabilities. *Aust. Jour. Statist.*, 17, 22-28.

[6] Cochran, W.G. (1946) Relative accuracy of systematic and stratified random samples for a certain class of populations. *Ann. Math. Statist.*, 17, 164-177.

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

[8] 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.

[9] 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.

[10] Midzuno, H. (1952) On the sampling system with probability proportional to sum of sizes. *Ann. Inst. Statist. Math.*, 3, 99-107.

[11] Mukhopadhyay, Parimal (1974) πPS sampling schemes to base HTE. *Cal. Statis. Assoc. Bull.*, 23, 20-24.

[12] Rao, J.N.K. (1963) On two systems of probability sampling schemes without replacement. *Ann. Inst. Statist. Math.*, 15, 67-73.

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

[14] Sen, A.R. (1953) On the estimation of the variance in sampling with varying probabilities. *Jour. Ind. Soc. Agri. Statist.*, 5, 119-127.