ON THE CHOICE OF A UNIQUE ESTIMATOR FOR THE IKEDA-SEN SAMPLING PROCEDURE*

BY

S.G. PRABHU-AJGAONKAR

Marathwada University, Aurangabad (MS)

(*Received in January, 1968*)

1. INTRODUCTION

Although the technique of drawing population elements with unequal probabilities was first introduced by Hansen and Hurwitz (1943), it received considerable impetus from Horvitz and Thompson (1952), who not only generalized this technique but also considered the problem of estimation.

It is noteworthy that the simplest way of incorporating the available supplementary information into a sampling procedure was envisaged by Ikeda, a student of Midzuno (1952) for Midzuno's estimator as follows :

On the first draw one population element is selected with unequal probabilities and on the second and subsequent draws, elements are selected with equal probabilities without replacement.

Sen (1951) discovered this procedure in connection with the Horvitz and Thompson estimator.

The problem under consideration is to provide a sample appraisal of the population total,

$$T_Y = \sum_{i=1}^{N} Y_i \qquad \ldots(1)$$

where Y_i denotes a measure of the characteristic under consideration for the ith population element, when a random sample of size n is

*Research Sponsored by Marathwada University, Aurangabad, India ; under Grant No. Research/12/68-69/3314-16.

drawn from a finite population of size N, in accordance with the Ikeda-Sen sampling procedure, utilising the ancillary information $(X_1, X_2, ..., X_N)$, where X_i represents a measure of the ancillary information on the ith population element.

$$X=\sum_{i=1}^{N} X_i \qquad ...(2)$$

and

$$p_i=\frac{X_i}{X},\ i=1, 2,..., N \qquad ...(3)$$

denote the probability set of the Ikeda-Sen sampling procedure. The different unbiased estimators that can be constructed for this purpose are described in the following section.

2. The Unbiased Estimators

A generalization of sampling from finite populations led Horvitz and Thompson (1952) to search for an appropriate estimator which in turn revealed the existence of three distinct classes of linear estimators. A unique estimator for their T_2 class of linear estimators derived by them is as follows :

$$t_2=\sum_{i}^{n} \frac{y_i}{\pi_i} \qquad ...(4)$$

where π_i $(i=1, 2,...,N)$ represents the probability of including the ith population element into the sample. For the Ikeda-Sen sampling procedure we have

$$\pi_i=\frac{N-n}{N-1}\, p_i+\frac{n-1}{N-1}\ . \qquad ...(5)$$

for $i=1, 2,...,N$.

The present author (1967) searched for a best estimator in Horvitz and Thompson's T_1 class of linear estimators when the Ikeda-Sen sampling procedure is adopted and discovered the estimator given by

$$t_1=y_1+\frac{(N-1)}{(n-1)}\sum_{r=2}^{n} y_r \qquad ...(6)$$

where y_r $(r=1, 2,..., n)$ denotes the outcome at the rth draw. Further he noted that

$$\text{var}\,(t_1)=\sum_{i=1}^{N} Y_i^2\frac{[(N-1)-p_i(N-n)]}{n-1}$$
$$+\sum_{i\neq j=1}^{N} Y_iY_j\,\frac{[(n-2)(N-1)+(N-n)(p_i+p_j)]}{(n-1)(N-2)}-T_Y^2 \qquad ...(7)$$

Horvitz and Thompson (1952) derived that

$$\text{Var }(t_2)=\sum_{i=1}^{N}\frac{Y_i^2}{\pi_i}+\sum_{i\neq j=1}^{N}\frac{Y_iY_j\pi_{ij}}{\pi_i\pi_j}-T_Y^2 \qquad \ldots(8)$$

where $\pi_{ij}(i\neq j=1, 2,\ldots, N)$ represents the joint probability of including the ith and jth population elements into the sample. For the Ikeda-Sen sampling procedure we have,

$$\pi_{ij}=\frac{(n-1)}{(N-1)}\left[\frac{(N-n)}{(N-2)}(p_i+p_j)+\frac{n-2}{N-2}\right] \qquad \ldots(9)$$

While presenting an outline of theory of sampling system, Midzuno (1950) discovered the estimator,

$$t_3=\frac{\Sigma y_i}{\Sigma p_i} \qquad \ldots(10)$$

where Σy_i is the sample total of the characteristic under consideration.

Midzuno showed that

$$\text{var }(t_3)=\sum_{i=1}^{N} Y_i^2 \sum_{s\supset i}\frac{1}{\binom{N-1}{n-1}\Sigma p_i}+\sum_{i\neq j=1}^{N} Y_iY_j \sum_{s\supset i,j}\frac{1}{\binom{N-1}{n-1}\Sigma p_i}-T_Y^2 \qquad \ldots(11)$$

where $\sum\limits_{s\supset i}$ denoted summation over those samples which include the ith population element. Similarly $\sum\limits_{s\supset i,j}$ stands for summation over those samples which contain the ith and jth population elements.

It is noted that there are three unbiased estimators belonging to the three classes of linear estimators formulated by Horvitz and Thompson (1952) when the Ikeda-Sen sampling procedure is adopted. Apparently, the problem of selecting a unique estimator is formidable one due to the fact that the usual principle of minimum variance does not render a solution.

4. Comparison of Estimators

We would like to resort to the criterion of the necessary best estimator, proposed by the present author (1965) for the selection of a unique estimator in the non-empty class for which a best estimator does not exist. Accordingly, we reproduce below the corrected version of the criterion.

Between unbiased estimators t and t' with variances

$$\text{var}\ (t) = \sum_i a_i Y_i^2 + \sum_{i \neq j} a_{ij} Y_i Y_j$$

and

$$\text{var}\ (t') = \sum_i b_i Y_i^2 + \sum_{i \neq j} b_{ij} Y_i Y_j$$

the estimator t is termed necessary better than t' if $b_i \geqslant a_i$ for all i with strict inequality for at least one i. In the original definition the equality sign was not there, which gave rise to the inaccuracy pointed out by Rao and Singh (1969).

Further, if an unbiased estimator in a class is a necessary better estimator than every other unbiased estimator of the class, that estimator is called a necessary best estimator of the class.

It is noted that the estimator t_2 is necessary better than the estimator t_3 if

$$\sum_{s \supset i} \frac{1}{\binom{N-1}{n-1} \Sigma p_i} > \frac{1}{\pi_i} \qquad \ldots(12)$$

for all i.

From the Cauchy-Schwarz inequality, We have

$$\left[\sum_{s \supset i} \frac{1}{\binom{N-1}{n-1} \Sigma p_i} \right] \left[\sum_{s \supset i} \frac{\Sigma p_i}{\binom{N-1}{n-1}} \right] > 1$$

where

$$\sum_{s \supset i} \frac{\Sigma p_i}{\binom{N-1}{n-1}} = \frac{\binom{N-1}{n-1} p_i + \binom{N-2}{n-2} (1-p_i)}{\binom{N-1}{n-1}}$$

$$= p_i + \frac{(n-1)}{(N-1)} (1-p_i)$$

$$= \frac{(N-n)p_i + n-1}{N-1}$$

$$= \pi_i \text{ from equation (5).}$$

Consequently

$$\sum_{s \supset i} \frac{1}{\binom{N-1}{n-1} \Sigma p_i} > \frac{1}{\pi_i} \quad \text{for all } i, \qquad \ldots(13)$$

and t_2 is necessary better than t_3 for the Ikeda-Sen sampling procedure. It is interesting to note that this result follows from that of Koop (1957) where he has compared the T_2 and T_3 classes of linear estimators for a general sampling procedure.

Further, the estimator t_2 is a necessary better estimator than t_1 if

$$\frac{[(N-1)-p_i(N-n)]}{(n-1)} > \frac{1}{\left[\frac{(N-n)}{(N-1)}p_i + \frac{n-1}{N-1}\right]}$$

for all i derived from equations at (5), (6) and (11).

This reduces to the inequality

$$p_i(1-p_i) > 0 \text{ for all } i,$$

which is true.

Consequently the estimator t_2 is a necessary better estimator than the estimator t_1 when the Ikeda-Sen sampling procedure is employed. Admittedly, the estimator t_2 is a necessary best among these three estimators.

As has already been pointed out that the technique of unequal probabilities is employed in order to incorporate the available ancillary information into the sampling procedure as a basis for determining the selection probabilities with which different population elements enter into the sample, reducing thereby the resulting sampling error of the estimator. Apparently the sampling procedure for which the sampling error is minimum will be selected for the given estimator. The problem of determining optimum sampling procedures for these estimators is dealt with in the next section.

5. Optimum Sampling Procedures

From the expression of var (t_2) in equation (6), it is noted that whenever

$$\pi_i = KY_i,\ i=1, 2, \ldots, N$$

it vanishes showing thereby that the sampling procedures possessing this property are optimum. In practice, Y_i's will not be known in advance. However, X_i's highly correlated with Y_i's and which constitute the available ancillary information are substituted for this purpose. Consequently, the optimum sampling procedures incorporating the available ancillary information are determined in practice as follows :

$$\pi_i = \frac{nX_i}{X},\quad i=1, 2, \ldots, N \qquad \ldots(14)$$

Apparently for the above relation to hold good, we should have $nX_i < X$, $i=1, 2, \ldots, N$. If for any X_i, $nX_i > X$ then the principle of stratification can be employed. In statistical literature many sampling procedures have been contrived to satisfy the above relation.

For the Ikeda-Sen sampling procedure, we have

$$\pi_i=\frac{(N-n)}{N-1}p_i+\frac{n-1}{N-1}\ ,\ i=1.\ 2,\ \ldots,\ N. \qquad \ldots(15)$$

Solving the equations at (14) and (15), we derive the optimum set of probabilities given by

$$p_i'=\frac{N-1}{N-n}\ \frac{nX_i}{X}-\frac{n-1}{N-n},\ i=1,\ 2,\ \ldots,\ N. \qquad \ldots(16)$$

Apparently the p'_i's are subject to the following two conditions ;

(1) $p'_i \geqslant 0,$

(2) $\sum_{i=1}^{N} p_i'=1,$

which give rise to the possibility of some of the p_i' being negative. In fact this occurs, when

$$\frac{nX_i}{X}<\frac{n-1}{N-1}\ . \qquad \ldots(17)$$

In such cases we can either follow the principle of stratification or attach the minimum value to p_i', namely, zero.

It is easily noted that the variance of t_3 vanishes whenever

$$\Sigma\, p_i'\ \alpha\ \Sigma\, Y_i$$

which for practical purposes reduces to

$$\Sigma\, p_i'\ \alpha\ \Sigma x_i. \qquad \ldots(18)$$

Apparently the optimum p'_i's are the same as given by the equation at (3).

For the minimization of var (t_1) with respect to p'_i's, we note that

$$\text{Min}\ [\text{var}\ (t_1)]=\sum_{i=1}^{N}\frac{(N-1)}{(n-1)}\,Y_i^2+\sum_{i\neq j=1}^{N}\frac{(n-2)(N-1)}{(n-1)(N-2)}Y_iY_j$$

$$-T_Y^2-\text{Max}\Bigg[\sum_{i=1}^{N}\frac{(N-n)}{(n-1)}p_iY_i^2$$

$$-\sum_{i\neq j=1}^{N}\frac{(N-n)(p_i+p_j)}{(n-1)(N-2)}Y_iY_j\Bigg] \qquad \ldots(19)$$

Subsequently it is noted that

$$\text{Min}\ [\text{var}\ (t)] = \text{Min}\left[\sum_{i=1}^{N} p_i Y_i(2T_Y - NY_i)\right] \qquad \ldots(20)$$

Let $Z_1, Z_2, \ldots, Z_N$ be the ranking when the N quantities $Y_i\,(2T_Y - NY_i)$ are ranked in the decreasing order of magnitude. Accordingly, we get

$$\text{Min}\ [\text{var}\ (t_1)] = \text{Min}\left[\sum_{i=1}^{N} P_i Z_i\right] \qquad \ldots(21)$$

The minimum occurs when P_N, corresponding to Z_N, is one and accordingly the rest of P_i's are zero. This is readily noted as follows :

$$\sum_{i=1}^{N} P_i Z_i = \sum_{i=1}^{N-1} P_i Z_i + \left(1 - \sum_{i=1}^{N-1} P_i\right) Z_N$$
$$= \sum_{i=1}^{N-1} P_i (Z_i - Z_N) + Z_N . \qquad \ldots(22)$$

Apparently the minimum occurs for the variation of P_i's when $P_i = 0$, for $i = 1, 2, \ldots, N-1$ and subsequently $P_N = 1$. Since the determination of P_i's depends on the knowledge about Y_i's as has already been explained, in practice X_i's are utilised. Let Y' be the value of Y_i corresponding to the X_i which is minimum among the N quantities $X_i(2T_X - NX_i)$. Accordingly, the minimum value of var (t_1) is noted to be

$$\text{Min}\ [\text{var}\ (t_1)] = Y'^2 + \frac{(N-1)}{(n-1)} \Sigma Y_i^2$$
$$+ \frac{(n-1)(N-1)+(N-n)}{(n-1)(N-2)}\, Y' \sum_{Y_i \neq (Y')} Y_i$$
$$+ \sum_{Y_i \neq Y_j \neq (Y')} Y_i Y_j \frac{(n-2)(N-1)}{(n-1)(N-2)} - (\Sigma Y_i)^2 \qquad \ldots(23)$$

A pair consisting of an unbiased estimator and a sampling procedure has been termed sampling—estimating strategy by Hajek (1959). Earlier, Midzuno (1950) denoted it by Sampling System. Since the term sampling system has also been used in the sense of a sampling procedure we would like to use the term sampling strategy in our discussion.

As in this section we have noted three sampling strategies for the Ikeda-Sen sampling procedure, the problem of selection of a unique sampling strategy crops up a new for which purposes we would like to follow the criterion of necessary better sampling strategy, formulated on the similar lines described in section 4. The next section is concerned with this.

6. Comparison of Sampling Strategies

Between two unbiased sampling strategies $H(t, p)$ and $H(t', p')$ with variances

$$\text{var}(t)=\sum_i a_i Y_i+\sum_{i\neq j} a_{ij} Y_i Y_j$$

and

$$\text{var}(t')=\sum_i b_i Y_i+\sum_{i\neq j} b_{ij} Y_i Y_j$$

the sampling strategy $H(t, p)$ is termed necessary better than $H(t', p')$ if

$b_i \geqslant a_i$ for all i with strict inequality for at least one i.

Moreover if an unbiased sampling strategy is a necessary better sampling strategy than every other unbiased sampling strategy, that sampling strategy is a necessary best sampling strategy.

When the Ikeda-Sen sampling procedure is adopted, the Midzuno sampling strategy is necessary better than the Horvitz and Thompson sampling strategy if,

$$\frac{X}{nX_i}>\sum_{s\supset i}\frac{1}{\binom{N-1}{n-1}}\frac{X}{\Sigma X_i}, \; i=1, 2, \ldots, N \quad \ldots(24)$$

It is noted that

$$\left[\sum_{s\supset i}\frac{1}{\binom{N-1}{n-1}}\frac{X}{\Sigma X_i}\right]\left[\sum_{s\supset i}\frac{\Sigma X_i}{\binom{N-1}{n-1}X}\right]\geqslant 1$$

i.e.,

$$\left[\sum_{s\supset i}\frac{1}{\binom{N-1}{n-1}}\frac{X}{\Sigma X_i}\right]\geqslant\frac{(N-1)X}{(n-1)X+X_i(N-n)} \quad \ldots(25)$$

From the inequalities (24) and (25), we derive

$$\frac{X}{nX_i}>\frac{(N-1)X}{(n-1)X+X_i(N-n)} \text{ for all } i. \quad \ldots(26)$$

This gives, $X>NX_i$, $i=1, 2, \ldots, N$. ...(27)

Since $X=\Sigma X_i$, we note that the above inequality cannot hold good. Consequently for the Ikeda-Sen Sampling procedure the Midzuno sampling strategy is not necessary better than the Horvitz and Thompson sampling strategy.

Next we proceed to examine whether the sampling strategy consisting of the estimator t_1 is necessary better than the Horvitz and Thompson sampling strategy. This is so if

$$\frac{nX'}{X}>0$$

for the population element Y' described in equation (23). Interestingly enough, the condition holds good. For all other population elements,

$$\frac{NX_i}{X}>\frac{N-n}{n-1}. \qquad \ldots(28)$$

Summing over i, we obtain

$$(n-1)n(X-X')>(N-1)(N-n)X$$

i.e.,

$$0>[(N-1)(N-n)-n(n-1)]X+n(n-1)X'$$

$$[N^2-Nn-N-n^2]X+n(n-1)X'<0,$$

which is not true.

Consequently we note that for the Ikeda-Sen sampling procedure, the sampling strategy containing the estimator t_1 is not necessary better than the Horvitz and Thompson sampling strategy and eventually that there does not exist a necessary best sampling strategy.

SUMMARY

In accordance with the Ikeda-Sen sampling procedure, a sample of size is drawn by selecting population units with unequal probabilities on the first draw and with equal probabilities without replacement on the subsequent $(n-1)$ draws. For this sampling procedure the present paper presents three unbiased estimators and attempts to select one with the help of the criterion of necessary best estimator suggested by the present author. Accordingly, it is noted that Horvitz and Thompson's estimator is a necessary best estimator for Ikeda-Sen's sampling procedure.

The well-known method of utilizing related information, that of incorporating it into the sampling procedure as a basis for determining the probabilities with which various population elements enter into the sample, is the technique of varying probabilities. Accordingly, the optimum sampling procedure corresponding to these three estimators are determined, which give rise to sampling

strategies. In accordance with the definition furnished by Hajek (1959), a sampling strategy is a pair of an estimator and a sampling procedure. Naturally, the problem of selecting a unique sampling strategy crops up, which, when the criterion of necessary best sampling strategy is applied, reveals that there does not exist a necessary best sampling strategy.

REFERENCES

1. Hajek, J. (1959) : 'Optimum strategy and other problems in probability sampling'. Casopis. Pest. Mat, 84, pp. 387-423.
2. Hansen, M.H., and Hurwitz, W.N. (1943) : 'On the theory of sampling from finite populations'. Ann. Math. Statist, Vol. 14, pp. 333-362.
3. Horvitz, D.G., and Thompson, D.J. (1952) : 'A generalization of sampling without replacement from a finite universe'. Jour. Amer. Statist. Assoc. Vol. 47, pp. 663-685.
4. Koop, J.C. (1957) : 'Contribution to the general theory of sampling finite populations without replacement and with unequal probabilities'. North Carolina, Institute of Statistics Mimeograph Series No. 296.
5. Midzuno, H. (1950) : 'An outline of the theory of sampling systems.' Ann. Inst. Statist. Math., Tokyo, Vol. 1, pp. 149-156.
6. Midzuno, H. (1952) : 'On the sampling system with probability proportionate to sum of sizes.' Ann. Inst. Statist. Math., Tokyo, Vol. 3, pp. 99-107.
7. Prabhu-Ajgaonkar, S.G. (1965) : 'On a class of linear estimators in sampling with varying probabilities without replacement'. Jour. Amer. Statist. Assoc., Vol. 60, pp. 637-642.
8. Prabhu-Ajgaonkar, S. G. (1967) : On Horvitz and Thompson's T_1 class of linear estimators, Ann. Math. Statist., Vol. 38, pp. 1882-1886.
9. Rao, J. N. K., and Singh, M. P. (1969) : 'On the Criteria of 'Hyper-admissibility' and 'necessary bestness' in survey Sampling. Tech. Report No. Math-Stat/40/69 Indian Statistical Institute, Calcutta.
10. Sen, A. R. (1951) : 'Present status of probability sampling and its use in the estimation Farm Characteristics'. Joint meeting of Econometric Society with Inst. Math. Stat., Minneapolis, Minn., (Abstract in the Jour. Econ. Soc., 1952, Vol. 20, pp. 103.