AN IPPS SAMPLING SCHEME USING LAHIRI'S METHOD OF SELECTION

By

K. Sankaranarayanan

Indian Statistical Institute, Calcutta

(Received in November, 1969)

1. Introduction

Consider a finite population consisting of N identifiable units denoted by the integers 1, 2,....., N. Let the value of the i-th unit for the variate under study, y, and an auxiliary variate, x be Y_i and X_i respectively. Let

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

and $$P_i = X_i/X.$$

For any sampling design defined by the set S of all possible samples, s, and the probability of selection $P(s)$ assigned for each sample s, such that each $P(s) > 0$ and $\sum_{s \in S} P(s) = 1$ let π_i denote the probability of unit i being included in the selected sample and π_{ij} denote that of units i and j being included in the selected sample. Then the Horvitz and Thompson estimator of Y, denoted by $\hat{Y}_{\mathrm{HT}}$ is given by

$$\hat{Y}_{\mathrm{HT}} = \sum_{i \in s} Y_i/\pi_i$$

It is well known that the efficiency of $\hat{Y}_{\mathrm{HT}}$ improves if π_i is made proportional to P_i ; *i.e.* if the inclusion probability is proportional to size (IPPS, for short).

Many writers have given sampling schemes which result in π_i proportional to P_i. Hanurav (1967), briefly reviews many such earlier schemes in the light of some desirable properties he suggests

for them in his paper in which he gives his own schemes for sample size 2. Rao and Bayless (1969) have mentioned about many such schemes while reporting on the rather extensive empirical studies they have carried out on various PPS without replacement strategies.

Apart from these, recently Sampford (1967) has suggested a new scheme which consists essentially in ensuring that the probability of selection of a 'particular sample', $S(n)$, consisting of n distinct units is

$$P \mid S(n) \mid = n\, K_n\, \lambda_1\lambda_2 \ldots \lambda_n \left(1 - \sum_{i=1}^{n} P_i\right)$$

where $\lambda_i = P_i/(1 - nP_i)$ and

$$K_n = \left\{ \sum_{t=1}^{n} \frac{t\, L_{n-t}}{n^t} \right\}^{-1}$$

where $L_0 = 1$ and $L_m = \sum_{S(m)} \lambda_1\lambda_2 \ldots \lambda_m$

where $\sum_{S(m)}$ denotes summation over all possible sets of m distinct units in the population.

In this paper a new sampling scheme is given which results in π_i's proportional to P_i's. This also consists in selecting samples of a fixed number of distinct units as a whole with pre-assigned probabilities. While all the previous schemes involve more or less heavy computational work, the present procedure is simple in concept and easily applicable in practice—both for selecting the sample and for evaluation of π_{ij}'s which are needed for estimating the variance. It has however a draw-back in that it is applicable only to populations having auxiliary information satisfying a certain condition.

The actual procedure is described in Section 2. In Section 3, the properties of the scheme vis-a-vis the estimation of the variance of $\hat{Y}_{HT}$ are discussed. In Section 4, results of a small empirical study, in which the present scheme is compared with certain similar strategies, are presented.

2. The Suggested Scheme

For our purpose, we consider a sample s as an unordered set consisting of a fixed number, n, of distinct units from the population

of N units. The set S (as defined in Section 1) of all possible samples consists of $\binom{N}{n}$ elements. Then the following result is easily proved :

Theorem : If the sth sample is selected with probability $P(s)$ proportional to $\sum_{j\in s}\left(P_j-\frac{1}{n}\frac{n-1}{N-1}\right)$ then π_i is proportional to P_i.

Proof : Let

$$P(s)=K\left[\sum_{j\in s}\left(P_j-\frac{1}{n}\frac{n-1}{N-1}\right)\right]$$

where K is a constant.

Then

$$\pi_i=\sum_{s\supset i}P(s)=K\sum_{s\supset i}\sum_{j\in s}P_j-\sum_{s\supset i}K\frac{n-1}{N-1}$$

$$=K\binom{N-1}{n-1}P_i+K\binom{N-2}{n-2}\sum_{j(\neq i)=1}^{N}P_j-\sum_{s\supset i}K\frac{n-1}{N-1}$$

because, there are $\binom{N-1}{n-1}$ samples including unit i and out of these there are $\binom{N-2}{n-2}$ samples including each unit j $[j(\neq i)=1, 2,..., N]$

$$=K\binom{N-1}{n-1}P_i+K\binom{N-2}{n-2}(1-P_i)-\binom{N-1}{n-1}K\frac{n-1}{N-1}$$

$$=\frac{K(N-2)!}{(N-n-1)!\,(n-1)!}P_i \qquad ...(1)$$

i.e., π_i is proportional to P_i ...*Q.E.D.*

The constant of proportionality K can be obtained by putting

$$\sum_{s\in S}P(s)=1$$

and is given by

$$K=\frac{n!\,(N-n-1)!}{(N-2)!}$$

Substituting the value of K in (1), we get $\pi_i=nP_i$.

For convenience in sampling, the selection procedure suggested by Lahiri (1951) for selection of samples with probability proportional

to total size may be used. The method can be described as follows. Let M be the maximum value of $\sum_{i\in s}\left(P_i-\frac{1}{n}\frac{n-1}{N-1}\right)$ for $s\in S$ select a simple random sample without replacement of size n. Let it be s_1. Find $\sum_{i\in s_1}\left(P_i-\frac{1}{n}\frac{n-1}{N-1}\right)=Q_1$ say. Select a random number R from 1 to M. If $R\leqslant Q_1$, s_1 is finally selected ; otherwise it is rejected and another sample s_2 is selected and this process is continued till a sample is finally selected.

The scheme can be applied if and only if $P(s)>0$ for all $s\in S$; it follows that the proposed scheme can be applied to all populations for which the ancillary characteristics satisfy the following condition—

$$\sum_{j\in s} P_j > \frac{n-1}{N-1} \text{ for all } s\in S \qquad \ldots(2)$$

(It may be seen that for applying Midzuno-Sen procedure, each P_i $(i=1,\ldots N)$ should be greater than $\frac{1}{n}\frac{n-1}{N-1}$; however, it is enough that each $P(s)$ is greater than $\frac{n-1}{N-1}$ for applying Lahiri's procedure). Verification of this condition for any given population is fairly easy, for, it is enough to see whether it is satisfied for the sample consisting of the n units having the smallest values of P_1.

3. Estimation of Variance

Let the condition (2) be satisfied, so that the scheme is applicable to the given population. Obviously all π_{ij}'s are positive and the actual values are evaluated as follows. By definition

$$\pi_{ij}=\sum_{s\supset i,j} P(s)=K\sum_{s\supset i,j}\left[\sum_{t\in s} P_t-\frac{n-1}{N-1}\right]$$

$$=\frac{n(n-1)}{N-2}\left(P_i+P_j-\frac{1}{N-1}\right) \qquad \ldots(3)$$

Now it is easy to obtain an unbiased estimate of the variance of $\hat{Y}_{HT}$ when $n\geqslant 2$.

Sen (1953) has shown that for the Midzuno-Sen scheme of selection in which $P'(s) \propto \sum_{i \varepsilon s} P_i$, $\pi_i'\pi_j' \geqslant \pi_{ij}'$ (where the symbols with prime denote the quantities for Midzuno-Sen scheme corresponding to those defined in Section 1) for all i and j $(i \neq j)$. It can be seen that if in this scheme, P_i is replaced by P_i' where

$$P_i' = \frac{P_i - \frac{1}{n}\frac{n-1}{N-1}}{\sum_{i=1}^{N}\left(P_i - \frac{1}{n}\frac{n-1}{N-1}\right)}$$

$$= \frac{n(N-1)}{N-n}\left[P_i - \frac{n-1}{n(N-1)}\right] \qquad \ldots(4)$$

the resultant value of $P'(s)$ equals $P(s)$ of the present scheme. For

$$P'(s) = \frac{1}{\binom{N-1}{n-1}} \sum_{i \varepsilon s} P_i'$$

$$= \frac{1}{\binom{N-1}{n-1}} \sum_{i \varepsilon s} \frac{n(N-1)}{(N-n)}\left[P_i - \frac{n-1}{n(N-1)}\right]$$

$$= \frac{n\,!\,(N-n-1)\,!}{(N-2)\,!}\left[\sum_{i \varepsilon s} P_i - \frac{n-1}{N-1}\right]$$

$$= P(s).$$

Therefore, the proposed scheme can be achieved by using Midzuno-Sen selection procedure replacing P_i by P_i' provided $P_i' > 0$ as by definition

$$\sum_{i=1}^{N} P_i' = 1$$

i.e.,
$$P_i > \frac{n-1}{n(N-1)} \quad \text{for all } i \qquad \ldots(5)$$

Incidentally, it may be noted that when (5) holds good $P_i < \frac{1}{n}$ for all i. Hence for the present scheme, $\pi_i\pi_j > \pi_{ij}$ for all i and j $(i \neq j)$ if (5) is satisfied. Thus under (5) the estimate of variance of $\hat{Y}_{\text{HT}}$ given by

$$v_{SYG}(\hat{Y}_{\text{HT}}) = \sum_{i \varepsilon s} \sum_{j(>i) \varepsilon s} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left\{\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right\}^2$$

and is always positive (see Sen, 1953).

4. Empirical Study

An empirical study has been carried out to compare the efficiencies of the following sampling strategies :—

(*A*) the proposed IPPS scheme with the estimator $\hat{Y}_{HT}$.

(*B*) the proposed IPPS scheme with the unbiased ratio estimator

$$\left(\sum_{i\in s} Y_i\right)/\left(\sum_{i\in s} P_i'\right)$$

where P_i' is given by (4).

(*C*) Lahiri's design, *i.e.*

$$P(s)\propto \sum_{i\in s} P_i \text{ with } \hat{Y}_{HT} \ ;$$

(*D*) Lahiri's design with the unbiased ratio estimator ;

$$\left(\sum_{i\in s} Y_i\right)/\left(\sum_{i\in s} P_i\right)$$

(*E*) Sampford's scheme of selection with $\hat{Y}_{HT}$

Taking $n=2$, the formulae for π_i and π_{ij} pertaining to the sampling schemes of the perceding strategies are given in Table (1).

Table (1) : Formulae for π_i and π_{ij} for the proposed, Lahiri and Sampford sampling schemes.

Sampling scheme	π_i	$\pi_{ij}=P(s) : s\supset i, j$
Proposed	$2P_i$	$\frac{2}{N-2}\left(P_i+P_j-\frac{1}{N-1}\right)$
Lahiri	$\frac{N-2}{N-1}P_i+\frac{1}{N-1}$	$\frac{(P_i+P_j)}{(N-1)}$
Sampford	$2P_i$	$2\left(1+\sum_{i=1}^{N}\frac{P_i}{1-2P_i}\right)^{-1}P_iP_j\left(\frac{1}{1-2P_i}+\frac{1}{1-2P_j}\right)$

If T_s is the value of a statistic T for sample s having selection probability $P(s)$, then the variance of T is given by

$$V(T)=E(T^2)-[E(T)]^2$$
$$=\sum_{s\in S} T_s^2P(s)-\left[\sum_{s\in S} T_sP(s)\right]^2$$

This was directly applied for getting all the variances in the empirical study which was carried out on IBM 1401 EDPM.

The population used was taken from Rao (1967) and is presented in Table (2).

Table (2) : Districts of Kerala State with 1961 Census population (y) and 1951 census population rounded to nearest thousand (x)

Serial No.	*Name of district*	X_i	Y_i	P_i
1.	Cannore	1375	1,780,294	0.10148
2.	Kozhikode	2065	2,617,189	0.15241
3.	Palghat	1565	1,776,566	0.11551
4.	Trichur	1363	1,639,862	0.10060
5.	Ernakulam	1530	1,859,913	0.11293
6.	Kottayam	1328	1,732,880	0.09801
7.	Alleppey	1521	1,811,252	0.11226
8.	Quilon	1474	1,941,228	0.10879
9.	Trivandrum	1328	1,744,531	0.09801

Since $N=9$, $n=2$, $(n-1)/n(N-1)=\cdot 06250$, it is now clear from col. 5,

Table (2) that condition (5) holds in this case.

The variance of the estimators of the population total Y were obtained for all the strategies under reference by enumerating all possible samples and computing their probabilities and the values of the estimates obtained from each sample. For strategies using $\hat{Y}_{HT}$ the variance of the Sen-Yates-Grundy estimator of variance of $\hat{Y}_{HT}$ were also calculated, after obtaining the value of this estimator for each possible sample using the values of π_i's and π_{ij}'s computed using the formulae given in Table (1). The results thus obtained are presented in Table (3).

Table (3) : Variance of estimators of total and Sen-Yates-Grundy variance estimator for strategies *A*, *B*, *C*, *D* and *E*

Strategy	*Variance of estimate of total*	*Variance of Sen-Yates-Grundy variance estimator*
A	$0.29749988 \times 10^{12}$	$0.10622146 \times 10^{24}$
B	$0.45025454 \times 10^{13}$	—
C	$0.86326588 \times 10^{12}$	$0.15840469 \times 10^{25}$
D	$0.29521540 \times 10^{12}$	—
E	$0.29753920 \times 10^{12}$	$0.10775736 \times 10^{24}$

It is seen from Table (3) that the strategies *B* and *C* are worse off compared to the rest ; *D* is slightly better than *A* and *E* while between *A* and *E*, *A* has a slight edge over *E* both in view of precision and stability of variance estimator as seen from cols. 2 and 3.

However, this empirical study is of limited scope and many more studies are needed to evaluate the comparative merits of various varying probability schemes available in literature. But it appears that the present scheme is likely to be as efficient as any other IPPS scheme in all cases where it is applicable.

Summary

For sampling with unequal probabilities without replacement, a scheme is given in this paper which yields inclusion probabilities proportional to size for all units in the population. This consists in selecting samples (of fixed number of distinct units) as a whole with pre-assigned probabilities, using Lahiri's method of selection. The scheme is applicable only if the sizes satisfy certain condition. It is also shown that the Sen-Yates-Grundy variance estimator of Horvitz-Thompson estimator of total is always positive under a slightly stricter condition. An empirical study, comparing the present scheme (with Horvitz-Thompson estimator) with similar strategies, is also included.

ACKNOWLEDGEMENTS

The author is thankful to Dr. M.N. Murthy for his valuable suggestions on the preparation of this paper. He is also grateful to the referee for his suggestions for improvement.

REFERENCES

1. Hanurav, T.V. (1967) : 'Optimum Utilization of Auxiliary Information', J.R.S.S. (B), Vol. 29, pp. 374-391.

2. Lahiri, D.B. (1951) : 'A Method of Sample Selection Providing Unbiased Ratio Estimates', Bull. Inter. Stat. Inst., Vol. 33, (2), pp. 133-140.

3. Rao, T.J. (1967) : Unpublished Ph.D. thesis submitted to the Indian Statistical Institute.

4. Rao, J.N K. and Bayless, D.L. (1969) : 'An Empirical Study of the Estimators and Variance Estimators in PPS Sampling', to appear in J.A.S.A.

5. Sampford, M.R. (1967) : 'On Sampling Without Replacement with Unequal Probabilities of Selection', Biometrika, Vol. 54, pp. 499-514.

6. Sen, A.R. (1953) : 'On the Estimate of Variance in Sampling with Varying Probabilities', Jour. Ind. Soc. Agr. Stat., Vol. 5, pp. 119-127.