A SAMPLING SCHEME TO REALIZE INCLUSION PROBABILITIES EXACTLY PROPORTIONAL TO SIZE

BY

S. SENGUPTA

Calcutta University, Calcutta

(Received : March, 1979)

SUMMARY

The familiar important problem of constructing a sampling design of fixed effective size with inclusion probabilities exactly proportional to size (IPPS, for short) is considered. A simple solution has been suggested. Further, the relative performance of the suggested procedure as compared to some other known procedures of similar type has been emperically investigated for a natural population.

INTRODUCTION

Consider a finite population of N identifiable units given by

$$U=(u_1, u_2, \ldots\ldots u_N)$$

Let Y and X be two real valued variates defined on U with $Y_i=Y(u_i)$ and $X_i=X(u_i)$. X_i's are called size measures and are known. Considering the problem of estimating the population total $T=\sum_{i=1}^{N} Y_i$ using the Horvitz Thompson (1952} estimator (HTE) we wish to select a sample of fixed effective size $n(\geqslant 2)$ such that the inclusion probability π_i of u_i is proportional to X_i *i.e.*

$$\pi_i=np_i \text{ for every } i=1, \ldots, N$$

where $p_i=X_i/X_1 \ldots\ldots+X_N$ to be called the normed size measure of u_i.

Although various solutions to this problem exist (for brevity we do not attempt to make a thorough review), an adequately satisfactory solution is yet to be achieved especially when the value of n exceeds two. In this note a procedure of such selection is suggested. The suggested scheme appears to be simple and may be looked upon as a simple generalization of simple random sampling without replacement (SRSWOR) procedure. An emperical study has also been made to judge the performance of the present procedure as compared to some other known procedures of similar type.

THE SUGGESTED PROCEDURE

Throughout the rest of this paper we take with no loss of generality

$$\frac{1}{n} > p_1 \geqslant p_2 \geqslant \ldots \geqslant p_N > 0.$$

A. Assume first $p_1 = p_n$. Let us define

$$\alpha_i = (n+i)(p_{n+i} - p_{n+i+1}) \; i = 0, 1, \ldots\ldots N_{-n}$$

with $$p_{N+1} = 0$$

Clearly, $$\alpha_i \geqslant 0 \; \forall i$$

and $$\sum_{i=0}^{N-n} \alpha_i = 1$$

The method then consists of the following steps.

I. Select one of the numbers 0, 1, ..., N_{-n} with probabilities $\alpha_0, \alpha_1, \ldots\ldots \alpha_{N-n}$. Let the selected number be *i*.

II. Take a SRSWOR sample of size *n* from

$$(u_1, u_2, \ldots\ldots, u_{n+i}).$$

Note: If $p_1 = p_N$ then $\alpha_{N-n} = 1$, so that the method reduces to SRSWOR procedure.

B. When $p_1 > p_n$ the same modifications as in Vijayan (1968) can be made as follows.

III. Choose one of the numbers 1, 2 ... *n* with probabilities $\delta_1, \delta_2, \ldots, \delta_n$ where,

$$\delta_r = n(p_{r-1} - p_r) \frac{S + (n-r+1)p_n}{S}, \; r = 1, 2, \ldots n$$

with $$S = p_{n+1} + p_{n+2} + \ldots + p_N$$

and $$p_o = 1/n.$$

Let the selected number be *r*.

VI. Take a sample of size $n-r+1$ from $(u_r, u_{r+1}, \ldots, u_N)$ by the procedure described in *A* with p_i's replaced by $p_i(r)$'s where

$$p_i(r) = \frac{p_n}{S + (n-r+1)p_n} \qquad i = r, r+1, \ldots, n$$

$$= \frac{p_i}{S + (n-r+1)p_n} \qquad i = n+1, \ldots, N$$

V. Add $u_1, \ldots, u_{r-1}$ to get the ultimate sample of *n* units.

INCLUSION PROBABILITIES OF FIRST TWO ORDERS

For the suggested procedure we have,

Theorem 1. Under scheme A, the inclusion probability for the i-th unit is

$$\pi_i = np_i \text{ for every } i=1 \ldots N$$

and the inclusion probabilities for the i-th and j-th unit together are given by

$$\pi_{ij} = np_n - p_{n+1} - n(n-1)\phi_{n+2}, \text{ for } 1 \leqslant i < j \leqslant n.$$

$$= n(n-1)\left[\frac{p\,j}{j-1} - \phi_{j+1}\right]$$

$$\text{for } i=1, 2, \ldots, j-1,$$

$$j=n+1, n+2, \ldots, N$$

where we define

$$\phi_j = \sum_{r=j}^{N} \frac{p_r}{(r-1)(r-2)}, \; j=n+1, \ldots, N \text{ and } \phi_{N+1}=0.$$

The proof follows directly from the description of scheme A and is omitted.

From the results of Vijayan (1968) it now follows;

Theorem 2. Under scheme B,

$$\pi_i = np_i \text{ for every } i=1, \ldots, N$$

$$\pi_{ij} = \sum_{r=1}^{n} \delta_r k_{ij}(r) \text{ for } 1 \leqslant i < j \leqslant N$$

where

$$k_{ij}(r) = 1, \qquad 1 \leqslant i < j \leqslant r-1$$

$$= \frac{(n-r+1)p_n}{S+(n-r+1)p_n}, \quad 1 \leqslant i \leqslant r-1, \; r \leqslant j \leqslant n.$$

$$= \frac{(n-r+1)p_j}{S+(n-r+1)p_n}, \quad 1 \leqslant i \leqslant r-1, \; n < j \leqslant N$$

$$= \pi_{ij}(r), \qquad r \leqslant i < j \leqslant N.$$

Here $\pi_{ij}(r)$ stands for the conditional inclusion probability of u_i and u_j in the sample of $(n-r+1)$ units obtained in step IV and may be calculated using the formulae given in theorem 1 with obvious modifications.

It is clear from the expressions of π_{ij}'s that they are non zero and are computable from simple and compact formulae so that unbiased variance estimates can be obtained without much labour.

A NUMERICAL ILLUSTRATIONS

To judge the performance of the suggested procedure we take up a practical example given by Sankarnarayanan (1969). The actual population is presented in table-1 and table-2 gives the actual variances of the estimator of the population total for samples of size two for the (IPPS, HTE) strategies corresponding to

(*i*) the present procedure

(*ii*) Shankarnarayanan's (1969) [or Midzuno's (1952)] procedure

(*iii*) Sampford's (1967) procedure.

TABLE 1

Districts of Kerala State with 1961 census population (Y) and 1951 census population rounded to nearest thousand (X)

Serial No.	*Name of the District*	Xi	Yi
1.	Cannore	1375	1,780,294
2.	Kozhikode	2065	2,617,189
3.	Palghat	1565	1,776,566
4.	Trichur	1363	1,639,862
5.	Ernakulam	1530	1,859,913
6.	Kottayam	1328	1,732,880
7.	Alleppey	1521	1,811,252
8.	Quilon	1474	1,941,228
9.	Trivendrum	1328	1,744,531

TABLE 2

Variance of the H.T.E. of population total for different sampling procedures

Sampling procedure	(*i*)	(*ii*)*	(*iii*)*
Variance of the HTE of the population total	.27310358 $\times 10^{12}$	.29749988 $\times 10^{12}$	.29753920 $\times 10^{12}$

Thus for the present example the variance of the HTE for the suggested scheme is considerably smaller than those for the other two IPPS procedures. But the study is too small to have any general conclusion.

*reproduced from Sankarnarayanan (1969).

Acknowledgement

The author is thankful to Sri P. Mallick of Calcutta University for making necessary computations presented in this paper and to the references for their helpful comments.

References

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

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

Sampford, M.R. (1967) : On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-514.

Sankarnarayanan, K. (1969) : An IPPS sampling scheme using Lahiri's method of selection. *Jour. Ind. Soc. Agri. Stat.*, 21(2), 58-66.

Vijayan, K. (1968) : An exact π PS sampling scheme—generalization of a method of Hanurav. *Jour. Roy. Stat. Soc.* (B), 30, 556-566.