A SAMPLING SCHEME WITH VARYING PROBABILITIES WITHOUT REPLACEMENT

BY

PADAM SINGH AND R. SINGH

Institute of Agricultural Research Statistics, New Delhi

(Received : November, 1975)

1. INTRODUCTION

1.1. Since a large unit, that is, a unit with a large value contribute more to the population total than smaller units, it is expected that sampling scheme which gives more chance of inclusion in the sample to larger units than to smaller units would provide more efficient estimators than those obtained from equal probability sampling schemes. Of all the sampling procedures with varying probabilities without replacement, pps systematic (ppss) sampling suggested by Medow [7] is commonly used, firstly because of its simplicity and secondly because it provides inclusion probability proportional to the size. Grundy [3], Hartley and Rao [5], Connor [1], Hartley [4] etc. have also contributed to the development of the ppss sampling.

1.2. One of the basic requirement of any sampling design is that it should provide an inbiased estimator of the sampling variance of the estimator of population parameter, but the ppss sampling does not provide us such an unbiased estimator for the variance of the estimator from the sample. In the present investigation a new sampling scheme with varying probability without replacement has been suggested which besides its simplicity, provides unbiased estimator of variance also. The relative efficiency of the new scheme as compared to that of the ppss sampling and pps with replacement (pps wr) sampling has been examined with the help of a numerical example.

2. THE NEW SAMPLING SCHEME

2.1. Suppose a finite population consists of N distinct and indentifiable units $U=(u_1, u_2 ......u_N)$ and a sample of size n is to be drawn from it. Let y and x be two real values characters defined over U which take values y_j and x_j respectively for the unit u_j,

$j=1, 2 \ldots N$ and assume that the vector $(x) = (x_1, x_2, \ldots, x_N)$ is known. We further assume that the characters y and x, known as study character and auxiliary character respectively are correlated with each other.

Now let n_1 and n_2 be some integers different from zero such that $n_1+n_2=n$. That the selection procedure under reference consists of the following steps.

A. Select a sample, say s_1, of size n_1 by ppss sampling, that is,

cumulate the sizes of units so that $Si=\sum_{1}^{i} x_j$

select a random number r from 1 to K:

where K is integer given by $\frac{S_N}{n_1}$:

(when $\frac{S_N}{n_1}$ is not an integer, the same can be made by properly multiplying x_i's by a constant).

Select the unit u_i in the sample if

$Si-1<r+jK\leqslant Si$, for any given $j=0, 1, 2, \ldots, (n_1-1)$.

B. Select a sample, say s_2, of size n_2 $(=n-n_1)$ from the remaining $(N-n_1)$ units of the population by simple random sampling without replacement (srswor).

The ppss sampling at 'A' provides inclusion probability of every unit proportional to its size in the sample s_1 and simple random sampling at 'B' provides equal chance for every unit to be included in the sample s_2.

2.2 *Estimate and Variance of Population Mean*

For the sampling scheme desribed in section 2.1, hereafter referred as modified ppss (mppss) sampling, Horvitz-Thompson (1952) estimator of the population mean can be worked out. For this the knowledge of inclusion probabilities for individual and pair wise units is required which can be obtained using theorems 2.2.1 and 2.2.2 presented below, without proof. These theorems are on the lines of Hartley (1966).

Theorem 2.2.1

The inclusion probability for unit u_i of population, in the sample of size n $(=n_1+n_2)$ by mppss sampling is given by

$$\pi_i=n_1\, p_i\, \frac{N-n}{N-n_1}+\frac{n-n_1}{N-n_1} \qquad \ldots(2.1)$$

where p_i is the probability of selection if *i-th* unit in the sample s_1.

Lemma 2.1

The inclusion probability π_{ij}' for a pair of units (u_i, u_j) in the sample s_1 is given by

$$\pi_{ij}' = \sum_{t=1}^{n_1-m} \pi_{ij}'(t)$$

where $\pi_{ij}'(t)$ denotes the inclusion probability for the unit u_j in the sample at *t-th* subsequent draw, when the unit u_i has already been included in the sample at some *m-th* draw and

(*i*) $\pi'_{ij}(t) = 0$ *if* $S_j - S_i - 1 \leq kt$

(*ii*) $= 0$ *if* $S_j - S_{i-1} > kt$

and $S_{j-1} - S_i \geqslant K t$

(*iii*) $=$ minimum of $\left[\frac{Kt - (S_j = 1 - S_i)}{K} \cdot \pi_j' \right]$

if $S_j - S_{i-1} > K t$

$S_{j-1} - S_i < K t$

and $S_{j-1} - S_{i-1} > K t$

(*iv*) $=$ minimum of $\left[\frac{S_j - S_{i-i} - K t}{K} \cdot \pi_i' \right]$

if $S_j - S_{i-1} > K t$

$S_{j-1} - S_i < K t$

and $S_{j-1} - S_{i-1} \leq K t.$

Theorem 2.2.2

Under the mppss selection procedure, the inclusion probability for a pair of units (u_i, u_j) in the sample of size $n(= n_1 + n_2)$ is given by

$$\pi_{ij} = \pi_{ij}' \left[1 + \frac{n_2 (n_2 - 1)}{(N - n_1)(N - n_1 - 1)} - \frac{2n_2}{N - n_1} \right] + \frac{(p_i + p_j) n_1 n_2 (N - n) + n_2 (n_2 - 1)}{(N - n_1)(N - n_1 - 1)}.$$

3. Main Feature of the Sampling Scheme :

One of the major advantages of the mppss sampling scheme discussed here is that it is possible to compute a set of revised probabilities of selection p_i' in the sample S_1 given by

$$p'_i = \frac{N - n_1}{n_1 (N - n)} \left\{ \frac{n x_i}{x} - \frac{n - n_1}{N - n_1} \right\}$$

such that the inclusion probability π_i resulting from the revised probabilities are proportional to x_i subject to

$$x_i > \frac{n-n_1}{n(N-n_1)} X \text{ for all } i \qquad \ldots(3.1)$$

The expression (3. 1) imposes a restriction on the selection procedure but the restriction is not severe.

4. Numerical Illustration

To illustrate the usefulness of new sampling scheme and to compare its efficiency with the usual pps systematic sampling and pps with replacement sampling, two populations have been studied, the data for which are given in table 4.1.

The data for population I have taken from studies by Horvitz Thompson [6] , relate to eye estimate of number of households (x)

TABLE 4.1

	Population-I			*Population-II*	
	x	y		x	y
1	18	19	1	120	126
2	9	9	2	290	257
3	14	17	3	70	108
4	12	14	4	141	185
5	24	21	5	497	421
6	25	22	6	130	148
7	23	27	7	87	119
8	24	35	8	198	338
9	17	20	9	357	368
10	14	15	10	151	221
11	18	18	11	83	121
12	40	37	12	153	151
13	12	12	13	156	224
14	30	47	14	327	410
15	27	27	15	304	275
16	26	25	16	139	177
17	21	25	17	623	790
18	9	13	18	150	176
19	19	19	19	217	236
20	12	12	20	91	129

and the actual number of households (y) in an area containing 20 blocks and for population II taken from the book *"The Design of Sample Surveys"* by Des Raj, relate to the, past population (x) and present population (y) of 20 towns of Greece.

The table 4.2 presents the variances for defferent sample sizes under different sampling schemes. From this table, it is of significance to note that Mitzuno-Sen [9] sampling scheme with revised probabilities, is not applicable for any sample size for either of the population considered here. Further it may be seen that the performance of mppss is almost always better than ppswr for both the populations and is better than ppss in most of the cases.

TABLE 4.2

n	n_1	n_2	*Variance of* mpps Population I	mpps Population II	ppss Population I	ppss Population II	ppswr Population I	ppswr Population II
3	2	1	3.20	—	7.32	644.54	5.41	783.38
4	3	1	5.53	363.12	3.60	618.35	4.05	587.53
5	4	1	1.62	323.37	3.23	335.96	3.24	470.03
6	4	2	1.29	—	1.06	340.83	2.70	390.69
	5	1	2.31	351·35	1.06	340.83	2.70	390.69
7	5	1	1.51	—	.65	280.51	2.31	335.73
	6	1	.91	185.40	.65	280.51	2.31	335.73
8	6	2	.71	—	.41	276.74	2.02	293.76
	7	1	.70	254.26	.41	276.74	2.02	293.76

SUMMARY

In this paper a new varying probability without replacement sampling scheme has been suggested. It is seen that the new scheme is not only very convenient and simple to operate but fulfils many requirements of an ideal sampling scheme. Firstly, it is possible to compute a set of revised probabilities of selection such that the inclusion probabilities are proportional to size measures for all the units. Secondly, the inclusion probability for every pair of units is positive which provides unbiased estimate of the sampling variance of the estimate of population mean. Further, the superiority of the suggested sampling scheme as compared to the usual pps systematic and pps with replacement sampling schemes has been examined empirically. Midzuno (1952) and usual pps systematic sampling scheme are particular cases of the new sampling scheme.

ACKNOWLEDGEMENT

The Authors are greatly thankful to the referees for their valuable comments in improving the paper.

REFERENCES

[1] Conner (1966) : An exact formula for the probability that the specified sample units will occur in a sample drawn with unequal probabilities and without replacement, JASA 384-396.

[2] Des Raj (1972) : Design of Sample Surveys, Mc Graw Hill Bcok Company.

[3] Grundy, P.M. (1954) : A method of sampling with probability exactly propprtional to size JRSS (B). **16**, 236-238.

[4] Hartiey, H.O. (1966) : Systematic sampling with unequal probabilities and without replacement. JASA 739-748.

[5] Hartley, H.O. and Rao, J.N.K. (1962) : Sampling with unequal probability without replacement. AMS.

[6] Horvitz, D.G. and Thompson, D.J. (1952) : A generalisation of sampling without replacement from a finite universe. JASA **47**, 663-685.

[7] Madow, W.G. (1949) : On the theory of systematic sampling X-II. AMS **20**, 333-354.

[8] Midzuno, H. (1952) : On the sampling system with porbability proportional to sum of sizes. AISM **3**, 99-107.

[9] Sen, A.R. (1953) : On the estimate of variance in sampling with varying probabilities. JISAS **5**, 119-127.

[10] Yates, F. and Grundy, P.M. (1953) : Selection without replacement from within strata with probability proporticnal to size. JRSS (B), **15**, 253-261.