IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 12, Issue 5 Ver. I (Sep. - Oct.2016), PP 01-04
www.iosrjournals.org

Alternative Sampling Strategy Based upon Coefficient of Mean Deviation When Auxiliary Information is Available

Spersh Bhatt
Department of Statistics, DSB Campus, Nainital,Uttarakhand, India

Abstract: *It is well know that in order to improve the efficiencies of the estimates probability sampling is preferred over non probability sampling. If the difference in the size of the units is large enough to affect the study, we make use of PPS sampling where the probability of selecting a unit is proportional to the size measure of the unit. Sometimes we may be confronted with situations where information on a character closely related to the main variable is available from a previous study or other secondary sources. Various authors have utilized this auxiliary information by taking the initial probability of selection equal to the size measure of the auxiliary information. This scheme however fails to give best results when the population under consideration is skewed. The paper presents an alternative without replacement sampling strategy obtained by utilizing auxiliary information to modify the initial probability of selection of the units. A computer program was developed in Visual Basic to find out the probabilities of selection and the variance of the sampling strategy proposed using the Horvitz Thompson estimator of population total. The empirical comparison of the proposed strategy with the existing Midzuno-Sen strategy shows that the proposed scheme performs better than the Midzuno-Sen strategy when the population is skewed.*

Keywords and Phrases: *coefficient of mean deviation, probability of selection, relative efficiency, sampling strategy, sampling design*

I. Introduction

It is well known that to avoid personal bias, random sampling is preferred over non random sampling. Attaching equal probability of selection to different units yields the method of simple random sampling. When unequal probabilities of selection are attached to different units in the population , it is called unequal probability sampling. If a sample from a finite population is drawn, usually the values of some character 'x' closely related to the main character of interest is available for all units of the population. The variable 'x' which is suitably normed, is often taken as a measure of the size of the unit. This occurs in socio-economic, agricultural and industrial surveys which are accompanied with the knowledge of past data. A unit with higher values of 'x' shall contribute more to the population total of main variable, than those with smaller sizes. One expects that, a selection procedure which gives higher selection probabilities to bigger units than to smaller units, should be more efficient than simple random sampling.

Consider a finite population 'U' of distinguishable units labeled 1,2,3,.........N. The collection of all possible samples is called the sample space denoted by 'S' .With each sample 's' a probability p(s) is attached which is the probability of drawing the sample 's' .

We thus have

(1) $$p(s) \geq 0$$

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

Here the sample from 'U' is an ordered sequence of labels from 'U' and represented by S = (i_1, i_2, ... , i_n)

where i_k is the label of the unit drawn at the k^{th} draw` and $1 \leq k \leq n$. The labels represent the units drawn with or without replacement in 'n' consecutive draws, hence the labels need not be distinct from each other. The size of the sample is 'n' and 'r' is the effective sample size(which is the number of distinct labels in 'S').

Let $\boldsymbol{P_i}$ denote the probability that the i^{th} unit is selected in the sample from the population .
By the addition law of probability

$$P_i = \sum_{i \in S} p(s)$$

where summation is taken over all possible samples containing the ith unit of the population. It is further assumed that p(s) is such that $\boldsymbol{P_i} > 0$ for
i = 1,2, ... ,N.

The collection S = {s} with a probability measure P = {p(s)}, defined on 'S', such that p(s) ≥ 0 and $\sum_{s \in S} p(s) = 1$ is called the sampling design and is denoted by D(S,P). A sampling procedure in which P_i (the probability of including the unit i in a sample of size n) is np_i. These are referred to as π-ps methods. Here p_i is the probability of selecting the ith unit of the population into the sample at the first draw. To estimate the population mean or total with such procedures, the commonly used estimator is the Horvitz-Thompson (H-T) estimator.

The unbiased H-T estimator for population total Y can also be written as

$$\hat{Y}_{HT} = \sum_{i=1}^{N} \frac{Y_i \delta_i}{P_i}$$

where $\delta_i = \begin{bmatrix} 1, & if \quad i \in S \\ 0 & otherwise \end{bmatrix}$

The variance of the H-T estimator for population total Y is given by

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{1-P_i}{P_i} Y_i^{\,2} + 2\sum_{i \langle j}^{N} \frac{(P_{ij} - P_i P_j)}{P_i P_j} Y_i Y_j$$

where i,j = 1, 2, 3, ...N.

Here P_{ij} is the probability of including the units i and j in the sample and

$$P_{ij} = \sum_{i,j \in s} p(s)$$

Yates and Grundy(1953) provided an alternative estimator of the population total Y , which is given by

$$V(\hat{Y}_{HT})_{YG} = \sum_{i \langle j}^{N} (P_i P_j - P_{ij}) \left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j} \right)^2$$

Some estimators of variance of the Horvitz-Thompson estimator have been given by Yates and Grundy and Sen (1953), Jessen(1969) and Ramakrishnan(1971).

Midzuno(1952)developed a sampling strategy in which the unit at the first draw is selected with unequal probability of selection. At all subsequent draws they are selected with equal probability and without replacement.

In the Midzuno-Sen scheme of probability proportional to size (pps) sampling the probability that the ith unit is included in the sample is given by

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

and the probability that both ith and jth units are included in the sample is given by

$$\frac{n-1}{N-1} \left[\frac{(N-n)}{(N-2)} (P_i + P_j) + \frac{n-2}{N-2} \right]$$

In the above scheme the probability of selection for a specific size ‘s’ is given by

$$\frac{\sum_{i \in s} x_i}{\sum_{s \in S} \left(\sum_{i \in s} x_i \right)}$$

Midzuno's scheme made π-ps has been considered by Rao(1963), Sankaranarayanan(1969), Chaudhuri(1974), Mukhopadhyay(1974) among others.

The Midzuno scheme though easy to implement is known to be less efficient in comparison to other unequal probability schemes. On the other hand Sampford scheme is known to be usually a good performer in the class of unequal probability schemes. This scheme however suffers from the drawback that it is rather difficult to implement particularly when n > 2.

Section 2 of the paper presents the methodology of obtaining the proposed strategy and the empirical study used for comparing the proposed strategy with the conventional Midzuno-Sen scheme.

Section 3 of the paper gives the tables and graphs giving the variance comparison of the Horvitz Thompson estimator under the proposed scheme and the Midzuno scheme.

II. Methodology and Empirical Study

Mean deviation for a sample of size ‘s’, for auxiliary information is given by

$$M.D. = \frac{1}{n}\sum_{i\in s}\left|X_i - A\right|$$

where ‘A’ is any measure of central tendency. It is now proposed that the average for the chapter on hand, for a sample of size ‘s’, is the mean given by $\overline{x}_s = \sum_{i\in s}\frac{X_i}{n}$.

The coefficient of mean deviation (about the mean),say CMD, is thus given as :

$$CMD = \frac{\sum_{i\in s}\frac{\left|X_i - \overline{x}_s\right|}{n}}{\overline{x}_s}.$$

It is now proposed that the probability of selection of a sample be given as

$$p(s) = \frac{CMD}{\sum_{s\in s}[CMD]} \qquad \text{…………}(\alpha)$$

It can be easily shown that the Horvitz-Thompson estimator under the above scheme is unbiased for the population total.

The empirical comparison for variance under the Midzuno scheme and the proposed one based on Coefficient of Mean Deviation has been done using a computer program developed in Visual Basic.

P_i and P_{ij} have been calculated on the basis of (α) and then the following Yates –Grundy formula for variance is used

$$V(\hat{Y}_{HT})_{YG} = \sum_{i\langle j}^{N}\left(P_iP_j - P_{ij}\right)\left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j}\right)^2$$

III. Tables for Efficiency Comparisons

To compare the two schemes 10 natural populations have been considered. These have been taken from Murthy (1977). Here Y stands for the number of cultivators in 1961 and X for area in 1951. Five cases have been considered for N=7 and n=3, two cases for N=8 and n=3 and three cases for N=9 and n=3. The first natural population for N=7 and n=3 along with the P_i and P_{ij} values have been given below.

Natural Populations for N=7, n=3

Natural Population 1

X	428	1177	1869	2544	2618	4113	4567
Y	193	819	611	806	1149	1510	1970

P_i and P_{ij} Values based on p(s) as given in (α)		
P_1 = 0.502401	P_{23} = 0.114344	P_{57} = 0 .096277
P_2 = 0.606957	P_{24} = 0 .039203	P_{67} = 0 .332695
P_3 = 0.332938	P_{25} = 0 .036538	
P_4 = 0 .249281	P_{26} = 0 .220125	
P_5 = 0 .236945	P_{27} = 0 .292918	
P_6 = 0 .545497	P_{34} = 0 .067813	
P_7 = 0 .672860	P_{35} = 0 .065903	
P_{12} = 0.217028	P_{36} = 0 .138614	
P_{13} = 0.071236	P_{37} = 0 .218331	
P_{14} = 0.075568	P_{45} = 0 .137901	
P_{15} = 0.080812	P_{46} = 0 .070657	
P_{16} = 0.262079	P_{47} = 0 .107420	
P_{17} = 0.298079	P_{56} = 0 .066824	

Further details regarding the natural populations and the P_i and P_{ij} values for these populations, as computed for the sampling design, using (α) maybe obtained from the author as the detailed description is not possible due to bravity.

Table 1: Description of Natural Populations for N=7, n=3

Sl. No.	Natural Population No.	Source	Variance M-S	Variance C.M.D.	% Relative efficiency of the estimator of C.M.D scheme over M-S scheme
1.	1	Murthy,(1977),Pg.127	7857190.04	3312670.51	237.19
2.	2	Ibid, Page 127	8321296.48	5157377.68	161.75
3.	3	Ibid, Page 127	3341035.17	1584939.51	210.80
4.	4	Ibid, Page 127	4038465.53	3217562.93	125.51
5.	5	Ibid, Page 127	8960843.40	5687162.82	157.56

Table 2: Description of Natural Populations for N=8, n=3

Sl. No.	Natural Population No.	Source	Variance M-S	Variance C.M.D	% Relative efficiency of the estimator of C.M.D scheme over M-S scheme
1.	1	Murthy, (1977), Pg. 129	1205843.96	991356.69	121.64
2.	2	Ibid, Page 129-130	1849844.39	1658772.73	111.52

Table 3: Description of Natural Populations for N=9, n=3

Sl. No.	Natural Population No.	Source	Variance M-S	Variance C.M.D	% Relative efficiency of the estimator of C.M.D scheme over M-S scheme
1.	1	Murthy,(1977),Pg.127	18053182.83	4937731.63	365.62
2.	2	Ibid, Page 127	24616710.38	7096773.58	346.87
3.	3	Ibid, Page 127	11123097.95	3226273.63	344.77

IV. Conclusion

From the efficiency comparisons given in the tables 1,2 and 3 above it can be concluded that the proposed strategy performs better than the existing strategy, specially in cases when the population is skewed.

References

[1]. Chaudhuri, A. (1974): On some properties of the sampling scheme due to Midzuno, Bull. Cal. Stat. Assoc., 18 ,1-24.

[2]. Horvitz, D.G. and Thompson, D.J. (1952): A generalisation of sampling without replacement from a finite universe. Jour. Amer. Stat. Assoc., 47, 663-85.

[3]. Jessen,R. J. (1969): Some methods of probability non-replacement sampling. Jour. Amer. Stat. Assoc.

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

[5]. Mukhopadhyay, P (1974): πps sampling schemes to base HTE. Cal. Stat. Assoc. Bull, 23 ,21-44.

[6]. Murthy,M. N. (1977): Sampling theory and Methods, 2nd ed., Statistical Publication SocietyCalcutta.

[7]. Ramakrishnan, M. K. (1971): Generalisation of Yates and Grundy estimates. Research Rep. 92/MKR/I, University of Sheffild, Manchester, Sheffild.

[8]. Rao, J. N. K. (1963): On three simple procedures of unequal probability sampling without replacement. Jour. Amer. Stat. Assoc., 59, 202-15.

[9]. Sampford,M.R.(1967): On sampling without replacement with unequal probabilities ofselection. Biometrika, 54, 499-513.

[10]. Sankarnarayanan,K(1969): An IPPS sampling scheme using Lahiri's method of selection. Jour. Ind. Soc. Agr. Stat., 21, 2, 59-66.

[11]. Sen,A.R.(1953): On the estimate of the variance in sampling with varying probabilities. Jour Ind Soc., Agr.Stat., 5, 119-27.

[12]. Yates,F. and Grundy,P.M.(1953): Selection without replacement within strata with probabilities proportional to size. Jour. Roy. Stat. Soc. B, 15, 243-261.