Brazilian Journal of Probability and Statistics (2007), 21, pp. 165–173.

On a πps scheme of sampling of two units

L. N. Sahoo,[1] **G. Mishra**[1] and **S. R. Nayak**[2]
[1] *Utkal University*
[2] *Dhenkanal College*

Abstract: The present paper proposes an inclusion probability proportional to size sampling scheme for sample of two units. This scheme possesses some desirable properties with regards to the inclusion probabilities, and provides an unbiased and non-negative variance estimator as is expected in the HT model. An empirical study with help of a wide variety of natural populations, is also undertaken to examine the performance of the suggested scheme compared to some other sampling schemes.

Key words: Inclusion probability, joint inclusion probability, πps sampling scheme.

1 Introduction

Let y_i and x_i, respectively, be the values of the study variable y and an auxiliary variable x (used as a size measure), for the ith unit of a finite population of N units with corresponding population totals $Y = \sum_{i=1}^{N} y_i$ and $X = \sum_{i=1}^{N} x_i$. Suppose that our aim is an estimation of Y based on a sample s of n units drawn from the population according to some unequal probability sampling without replacement scheme with π_i as the inclusion probability of ith unit, and π_{ij} as the joint inclusion probability of ith and jth units. The most commonly used estimator in this situation is the Horvitz-Thomson (1952) (HT) estimator defined by

$$\widehat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}.$$

From the general theory developed by Horvitz and Thomson (1952), we have $\sum_{i=1}^{N} \pi_i = n$, $\sum_{j \neq i=1}^{N} \pi_{ij} = (n-1)\pi_i$ and $\sum_{i=1}^{N} \sum_{j<i} \pi_{ij} = \frac{1}{2}n(n-1)$. An unbiased estimator of $\mathrm{Var}\,(\widehat{Y}_{HT})$, as suggested by Yates and Grundy (1953), is given by

$$\nu(\widehat{Y}_{HT}) = \sum_{i} \sum_{j \in s} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2. \tag{1.1}$$

A sufficient condition for (1.1) to be always non-negative is that $\pi_{ij} < \pi_i \pi_j$, $i \neq j$.

It is a well known result that considerable reduction in the variance of $\widehat{Y}_{HT}$ can be expected if π_i's are proportional to x_i. Such schemes are known as πps or IPPS (inclusion probability proportional to size) schemes. The estimator commonly used to estimate population mean or total with such schemes is the HT estimator. A number of πps schemes are available in the literature [cf., Brewer and Hanif (1983), Chaudhuri and Vos (1988)]. But, for the majority of these schemes, calculations of π_{ij} and expression for $\nu(\widehat{Y}_{HT})$ rapidly becomes cumbersome as $n > 2$. So, for simplicity many πps methods are restricted to $n = 2$ only. These methods have their application in stratified sampling, where stratification is sufficient deep i.e., the number of strata (and their sizes) is such that a sample of 2 units per stratum meets the requirement on the total sample size.

In this paper, we introduce a new πps sampling scheme for $n = 2$, having some desirable properties in terms of π_i and π_{ij}. The suggested scheme also performs well as compared to some popular sampling schemes for a number of natural populations.

2 Description of the suggested scheme

For the N units of the population, let us consider the set of revised probabilities $\{P_1, P_2, \ldots, P_N\}$, where P_i is defined by

$$P_i = \frac{(2p_i - \lambda h_i)(1 - h_i)}{1 - 2h_i}, \quad i = 1, 2, \ldots, N, \tag{2.1}$$

such that $p_i = x_i/X$ is the initial probability of selection of ith unit, $h_i = p_i(1 - p_i)\Big/\sum_{j=1}^{N} p_j(1 - p_j)$ and $\lambda = \sum_{i=1}^{N}(p_i/(1 - 2h_i))\Big/\sum_{i=1}^{N}(h_i(1 - h_i)/(1 - 2h_i))$. The constant λ is determined so as to make $\sum_{i=1}^{N} P_i = 1$, i.e., by solving the equation

$$2\sum_{i=1}^{N}\frac{p_i(1 - h_i)}{1 - 2h_i} - \lambda\sum_{i=1}^{N}\frac{h_i(1 - h_i)}{1 - 2h_i} = 1, \text{ for } \lambda. \tag{2.2}$$

It may be noted here that computation of the revised probabilities is restricted only to those situations for which $h_i < 1/2$ and $h_i < 2p_i/\lambda$ i.e., $h_i < \min(1/2, 2p_i/\lambda)$. These restrictions on h_i seem to be very much severe. But, our experiment with the help of a number of artificial and natural populations available in various text books as well as research papers on survey sampling confirms that they can meet for many practical situations.

Our suggested sampling scheme for $n = 2$ consists of the following steps:

Step I: Select the first unit, say i, with revised probability P_i and without replacement;

Step II: Select the second unit, say j, from the remaining $(N - 1)$ units with conditional probability

$$P_{j|i} = \frac{h_j}{1 - h_i}. \tag{2.3}$$

3 Inclusion probabilities and properties of the scheme

By definition,

$$\begin{aligned}\pi_i &= P_i + \sum_{j\neq i} P_j \frac{h_i}{1-h_j} \\ &= 2p_i - h_i\left[\lambda - 2\sum_{j=1}^{N}\frac{p_j}{1-2h_j} + \lambda\sum_{j=1}^{N}\frac{h_j}{1-2h_j}\right]. \qquad (3.1)\end{aligned}$$

Again from (2.2), on simplification, we also have

$$\lambda - 2\sum_{i=1}^{N}\frac{p_i}{1-2h_i} + \lambda\sum_{i=1}^{N}\frac{h_i}{1-2h_i} = 0. \qquad (3.2)$$

Hence, from (3.1) and (3.2) we obtain

$$\pi_i = 2p_i. \qquad (3.3)$$

The second order inclusion probabilities are

$$\pi_{ij} = P_i P_{j|i} + P_j P_{i|j} = \frac{(2p_i - \lambda h_i)h_j}{1-2h_i} + \frac{(2p_j - \lambda h_j)h_i}{1-2h_j}\ . \qquad (3.4)$$

The desirable properties of the suggested scheme are as follows:

(i) $\sum_{i=1}^{N}\pi_i = 2\sum_{i=1}^{N}p_i = 2$;

(ii) $\sum_{j\neq i}^{N}\pi_{ij} = \frac{2p_i - \lambda h_i}{1-2h_i}\sum_{j\neq i}^{N}h_j + h_i\sum_{j\neq i}^{N}\frac{2p_j-\lambda h_j}{1-2h_j}$
$= 2p_i - h_i\left[\lambda - 2\sum_{j=1}^{N}\frac{p_j}{1-2h_j} + \lambda\sum_{j=1}^{N}\frac{h_j}{1-2h_j}\right]$
$= 2p_i = \pi_i$. [using (3.2]

(iii) $\sum_{i=1}^{N}\sum_{j<i}\pi_{ij} = \frac{1}{2}\sum_{i\neq j}^{N}\pi_{ij} = 1.$

(iv) Proceeding in an obvious way as is given in Konijn (1973, p.253), for any arbitrary i and j, we obtain

$$\begin{aligned}\pi_i\pi_j - \pi_{ij} &= \frac{(2p_i - \lambda h_i)(2p_j - \lambda h_j)}{(1-2h_i)(1-2h_j)}\left(\sum_{k>2} h_k\right)^2 \\ &+ h_i h_j\left[\sum_{k>2}\frac{2p_k - \lambda h_k}{1-2h_k}\right]^2 + \pi_{ij}\sum_{k>2}\frac{(2p_k - \lambda h_k)h_k}{1-2h_k} \geq 0.\end{aligned}$$

Hence, the Yates-Grundy variance estimator of the HT estimator under the suggested sampling scheme is always non-negative.

4 Numerical study of the performance of the scheme

To study the performance of the proposed sampling scheme compared to some other well known sampling procedures, we consider two different performance measures, *viz.*,

(i) Efficiency with respect to probability proportional to size with replacement (PPSWR) scheme, and

(ii) Stability of the variance estimator.

Here, we accept Hanurav's (1967) criterion $\phi = \min(\pi_{ij}/(\pi_i\pi_j)) > \beta$, for β sufficiently away from zero, to study stability of the variance estimator.

The following eight sampling procedures are taken into consideration:

S_{WR}: Conventional estimator under PPSWR sampling scheme;

A: HT estimator under the sampling scheme of Brewer (1963);

B: HT estimator under the sampling scheme of Singh (1978);

C: HT estimator under the sampling scheme of Deshpande and Prabhu Ajgaonkar (1982);

S_{DR}: Ordered estimator of Raj (1956);

S_{MR}: Unordered estimator of Murthy (1957);

S_{RHC}: Estimator of Rao, Hartley and Cochran (1962);

S: HT estimator under the suggested sampling scheme.

Three πps sampling schemes A, B and C are considered for comparison in respect of efficiency and stability of the variance estimator. Because (i) these schemes are relatively simple to operate, (ii) they do not involve much mathematical complexity, and (iii) computation of π_{ij} for these schemes is also simple.

We have not included πps methods of Rao (1965), Durbin (1967) and Sampford (1967), because they give the same π_i and π_{ij} values which are identical to that of Brewer's methods. To examine the efficiency of HT estimator based on the suggested sampling scheme over other estimators based on probability proportional to size without replacement (PPSWOR) sampling scheme, we also include three well known estimators of Raj, Murthy and Rao-Hartley-Cochran in our comparison. Since a theoretical comparison is impracticable, we resort to an empirical study with the help of 20 natural populations.

Table 1 *Description of populations*

Pop.	Source	N	y	x	ρ
1	Singh and Singh Mangat (1996, p.193)	24	no. of dwellings occupied by tenants	no. of dwellings	0.85
2	Cochran (1977, p.203)	10	actual weight of peaches	estimated weight of peaches	0.97
3	Sukhatme and Sukhatme (1970, p.166, 1-10)	10	no. of banana bunches	no. of banana pits	0.65
4	Sukhatme and Sukhatme (1970, p.166, 11-14)	10	no. of banana bunches	no. of banana pits	0.84
5	Cochran (1977, p.187)	18	population in 1960	population in 1950	0.96
6	Horvitz and Thompson (1952)	20	no. of households	eye estimated no. of households	0.87
7	Singh and Singh Mangat (1996, p.199)	12	blood pressure	age	0.75
8	Cochran (1977, p.325)	10	no. of persons	no. of rooms	0.65
9	Raj and Chandhok (1998, p.291, 1-10)	10	actual no. of households	eye estimated no. of households	0.84
10	Raj and Chandhok (1998, p.291, 11-20)	10	actual no. of households	eye estimated no. of households	0.87
11	Mukhopadhyay (1998, p.131, 1-10)	10	population in 1971	population in 1961	0.99
12	Mukhopadhyay (1998, p.131, 11-20)	10	population in 1971	population in 1961	0.93
13	Singh and Singh Mangat (1996, p.79)	14	pet animals	households	0.98
14	Asok and Sukhatme (1976, 1-17)	17	acreage under oats in 1957	recorded acreage of crops and grass for 1947	0.39
15	Asok and Sukhatme (1976, 18-35)	18	acreage under oats in 1957	recorded acreage of crops and grass for 1947	0.61
16	Murthy (1967, p.400, sub-sample I)	10	current population	previous census population	0.98
17	Murthy (1967, p.400, sub-sample II)	10	current population	previous census population	0.97
18	Sukhatme and Sukhatme (1970, p.51, 1-13)	13	area under rice	total cultivated area	0.95
19	Sukhatme and Sukhatme (1970, p.51, 14-25)	12	area under rice	total cultivated area	0.98
20	Singh and Singh Mangat (1996, p.88)	18	total yield	area under wheat	0.99

Table 1 describes source, size (N), nature of y and x, and correlation coefficient between y and $x(\rho)$ of the populations under consideration. Numerical values of the relative efficiency of the comparable sampling procedures *w.r.t.* S_{WR} (in %), and stability parameter ϕ of variance estimators of the schemes A, B, C and S

are presented in Tables 2 and 3, respectively. Our calculations are based on all $C(N, n)$ possible samples of $n = 2$ drawn from a population. The entries for the most efficient and most stable variance estimator cases for each population are boldly printed.

Findings in Table 2 indicate that the suggested sampling procedure S is more efficient than A, B and C for all populations and more efficient than S_{DR}, S_{MR} and S_{RHC} for 17 populations. Relative efficiencies of IPPS schemes including S in comparison to PPSWOR methods are low for populations 17, 19 and 20 even if ρ values are extremely high. The reason is that the population regression line of y on x intercepts the y-axis at some distance from the origin.

Table 2 *Relative efficiency of different sampling procedures*

Pop.	Sampling Procedures							
	S_{WR}	A	B	C	S_{DR}	S_{MR}	S_{RHC}	S
1	100.00	104.867	104.890	104.667	104.726	104.972	104.545	**105.864**
2	100.00	112.109	112.094	112.119	111.123	112.534	112.500	**112.917**
3	100.00	113.740	113.763	113.721	111.663	113.220	112.709	**113.810**
4	100.00	112.304	112.384	112.315	111.252	112.711	112.539	**112.863**
5	100.00	107.079	107.114	107.065	107.366	107.993	106.250	**108.073**
6	100.00	107.844	107.921	107.841	106.607	107.101	105.555	**107.998**
7	100.00	110.919	110.360	110.918	109.602	110.635	110.000	**111.050**
8	100.00	111.655	111.668	111.656	110.902	112.249	112.300	**112.652**
9	100.00	118.321	118.531	118.322	113.586	115.773	113.231	**118.960**
10	100.00	118.524	118.917	118.523	114.131	116.619	115.677	**119.397**
11	100.00	101.662	100.006	101.665	109.626	110.406	110.456	**111.234**
12	100.00	113.075	113.346	113.076	113.556	115.928	113.202	**116.021**
13	100.00	115.758	112.926	115.756	112.158	114.271	108.333	**116.280**
14	100.00	108.015	108.041	108.005	106.927	107.454	106.250	**108.999**
15	100.00	106.975	106.862	106.976	105.610	106.097	106.250	**106.996**
16	100.00	111.005	110.971	111.001	111.816	113.567	112.501	**113.829**
17	100.00	115.851	116.583	115.849	116.684	**116.872**	116.717	116.641
18	100.00	114.115	114.103	114.114	111.226	112.815	108.333	**114.970**
19	100.00	109.576	109.909	109.575	111.615	**113.050**	110.059	110.491
20	100.00	104.834	104.739	104.832	106.415	106.714	**106.950**	104.859

Variance estimator of S is more stable than those of A, B and C for 16 populations (Table 3). It's low stability for populations 5, 9, 15 and 19 is probably because of the disproportionate variation of x-values to make the ratio $\frac{\pi_{ij}}{\pi_i \pi_j}$ very small for some samples. However, choice of other criterion [*cf.*, Rao and Bayless (1969)] may improve this stability a bit.

Table 3 *Stability parameter of different sampling schemes*

Pop.	Sampling Schemes			
	A	B	C	S
1	0.5312	0.5244	0.5313	**0.5320**
2	0.5440	0.5439	0.5441	**0.5449**
3	0.5421	0.5410	0.5420	**0.5425**
4	0.5359	0.5379	0.5349	**0.5455**
5	0.4655	**0.5002**	0.4658	0.4572
6	0.5039	0.5085	0.5037	**0.5089**
7	0.5340	0.5346	0.5337	**0.5365**
8	0.5477	0.5475	0.5475	**0.5486**
9	0.5166	**0.5219**	0.5164	0.5149
10	0.5034	0.5176	0.5035	**0.5192**
11	0.5443	0.5292	0.5441	**0.5503**
12	0.4658	0.5067	0.4656	**0.5116**
13	0.4845	0.4763	0.4824	**0.4868**
14	0.5102	0.5142	0.5103	**0.5294**
15	0.4926	**0.5054**	0.4928	0.4865
16	0.4950	0.5163	0.4491	**0.5374**
17	0.4915	0.5139	0.4917	**0.5283**
18	0.5010	0.5075	0.5008	**0.5093**
19	0.4399	**0.5007**	0.4397	0.4176
20	0.4976	0.5073	0.4976	**0.5157**

5 Conclusions

On the basis of the analytical and empirical results derived in this work, we may conclude that the suggested sampling procedure is no way inferior to some standard sampling procedures. But, no general conclusion can be drawn from the empirical study as the conclusion is based on the results for 20 populations only and the gain in efficiency of the suggested scheme compared to other leading alternatives is in fact rather small. However, this comparison gives an indication that the suggested scheme (if it exists) compares well with other popularized schemes in terms of efficiency as well as stability of the estimated variance.

Acknowledgements

The authors are grateful to the referee whose constructive comments led to an improvement in the paper.

(Received April, 2006. Accepted November, 2006.)

References

Asok, C. and Sukhatme, B. V. (1976). On Sampford's procedure of unequal probability sampling without replacement. *Journal of the American Statistical Association*, **71**, 912–918.

Brewer, K. R. W. (1963). Ratio estimation in finite populations: Some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, **5**, 93–105.

Brewer, K. R. W. and Hanif, M. (1983). Sampling with unequal probabilities. *Lecture Notes in Statistics*. Springer-Verlag.

Chaudhuri, A. and Vos, J. W. E. (1988). *Unified Theory and Strategies of Survey Sampling*. North Holland.

Cochran, W. G. (1977). *Sampling Techniques*. 3.rd. Edition. John Wiley and Sons.

Deshpande, M. N. and Prabhu Ajgaonkar, S. G. (1982). An IPPS (inclusion probability proportional to size) sampling scheme. *Statistica Neerlandica*, **36**, 209–212.

Durbin, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Applied Statistics*, **16**, 152–164.

Hanurav, T. V. (1967). Optimum utilization of auxiliary information: πps sampling of two units from a stratum. *Journal of the Royal Statistical Society B*, **29**, 374–391.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Konijn, H. S. (1973). *Statistical Theory of Sample Survey Design and Analysis*. North Holland.

Mukhopadhyaya, P. (1998). *Theory and Methods of Survey Sampling*. New Delhi: Prentice-Hall of India.

Murthy, M. N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhyã*, **18**, 379–390.

Murthy, M. N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.

Raj, D. (1956). Some estimators in sampling with varying probabilities without replacement. *Journal of the American Statistical Association*, **51**, 269–284.

Raj, D. and Chandok, P. (1998). *Sample Survey Theory*. Narosa Publishing House.

Rao, J. N. K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, **3**, 173–180.

Rao, J. N. K. and Bayless, D. L. (1969). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units per stratum. *Journal of the American Statistical Association*, **64**, 540–549.

Rao, J. N. K., Hartley, H. O. and Cochran, W. G. (1962). A simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society B*, **24**, 482–491.

Sampford, M. R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, **54**, 499–513.

Singh, P. (1978). The selection of samples of two units with inclusion probabilities proportional to size. *Biometrika*, **65**, 450–454.

Singh, R. and Singh Mangat, N. (1996). *Elements of Survey Sampling*. The Netherlands: Kluwer Academic Publishers.

Sukhatme, P. V. and Sukhatme, B. V. (1970). *Sampling Theory of Surveys with Applications*. Calcutta: Asia Publishing House.

Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, **15**, 235–261.

L. N. Sahoo and **G. Mishra**
Department of Statistics
Utkal University
Bhubaneswar 751004, India.
E-mail: lnsahoostatuu@rediffmail.com

S. R. Nayak
Department of Statistics,
Dhenkanal College, Dhenkanal
India.
E-mail: sony@uwlooemail